# LEARNING FROM A SERVICE GUARANTEE QUASI-EXPERIMENT

Xinlei (Jack) Chen
Souder School of Business
University of British Columbia

George John*
Pillsbury-Gerot Chair of Marketing
Carlson School of Management

Julie M. Hays
Assistant Professor
University of St. Thomas

Arthur V. Hill
John and Nancy Lindahl Professor
Carlson School of Management

Susan E. Geurs
Vice President, Carlson Hotels Worldwide

Edited June 6, 2008 forthcoming Journal of Marketing Research November 2009

* Corresponding author: Professor George John, Marketing Department, Carlson School of Management, University of Minnesota, Marketing and Logistics Department, 321 19th Avenue South, Minneapolis, MN 55455-0413, USA. Phone (612) 624-6841, Fax 612/626-8328, E-mail: johnx001@umn.edu. We gratefully acknowledge the financial support of the National Science Foundation and of the Carlson Companies for this research.

# LEARNING FROM A SERVICE GUARANTEE QUASI-EXPERIMENT


## ABSTRACT

We analyze data from a service guarantee program implemented by a mid-priced hotel chain. Using a multi-site regression discontinuity quasi-experimental design developed for these data from 85,321 guests and 81 hotels over 16 months, we control for unobserved heterogeneity among guests and treatments across hotels, and develop Bayesian posterior estimates of the varying program effect for each hotel. Our results contribute to theory and practice. First, we provide new insights into how service guarantee programs operate in the field. Specifically, we find that the service guarantee was more effective at hotels with a better prior service history and an easier-to-serve guest population, both of which are consistent with signaling arguments, but do not comport with the incentive argument that guarantees actually improve service quality. Second, our study offers managers better decision rules. Specifically, we devise program continuation rules that are sensitive to both observed and unobserved differences across sites. In addition, we devise policies to reward hotels for exceeding site-specific expectations. By controlling for observed and unobserved differences across sites, these policies potentially reward even sites with negative net program effects, which is useful in reducing the organizational stigma of failure. Finally, we identify sites that should be targeted for future program rollout by computing their odds of succeeding.

# INTRODUCTION

A number of studies have shown that service quality improvements represent a significant opportunity to improve customer satisfaction and firm profits (e.g., Anderson and Sullivan 1993; Anderson, Fornell and Lehmann, 1994; Fornell 1992; Hauser, Simester, and Wernerfelt 1994, 1996, 1997). According to marketing orthodoxy, a properly designed and implemented service guarantee (SG) can be an important tool for improving customer service evaluations. A SG is a particular type of warranty (Boulding and Kirmani, 1993) that promises a particular level of service to a customer and also promises compensation if that level of service is not achieved. SGs vary substantially both in the promised level of service (e.g., unconditional) and in the type of compensation (e.g., money back, free service next time). Spurred by the influential work of Hart (1988) and others, SGs have attracted considerable attention from industry. However, the limitations of current analysis tools handicap firms that seek to learn from their implemented programs. In particular, identifying where, when and why a SG works, targeting the best prospective sites, and rewarding good performers are among the keys to making financially responsible and accountable marketing program decisions as advocated by Rust *et al.* (1995). These decisions rest on a proper evaluation of the intervention, but current evaluation methods face two important and inter-related challenges – unobserved subject and treatment heterogeneity.

Unobserved Subject Heterogeneity: Field interventions invariably assign intact groups of customers to the treatment in contrast to laboratory designs that assign individual subjects. Shadish, Cook and Campbell (2002) document various analysis and inference problems with intact group interventions because of unobserved group differences. For example, Bolton and Drew's (1991) analysis of a field intervention discovered program effects at the individual

customer level, but not at the site level (at which the intervention occurred) which they attributed to unobserved subject heterogeneity across sites. This problem cannot be simply remedied by resorting to designs with random assignment at the individual level (see Hutchinson *et al*., 2000). Our challenge is to develop analysis tools that accommodate the unobserved subject heterogeneity endemic to these intact group quasi-experimental designs.

Unobserved Treatment Heterogeneity: Field interventions are invariably implemented by different managers at different locations at different time-points. As such, the conventional assumption of a constant causal effect of the intervention is strained, and it is more useful to think of a varying causal effect. Simester *et al*. (2000) provide the only instance of an analysis of a field quasi-experiment in marketing that accounts for treatment heterogeneity. However, their approach is specific to two-wave designs. Other quasi-experimental designs leave unanswered important managerial and policy questions such as the identification of those specific sites where the program should be continued or discontinued, and the design of tailored rewards for exceeding expected treatment effects. Our challenge is to develop tools for addressing such questions for field quasi-experiments beyond two-wave designs.

*Goals of paper*

The goals of this paper are to learn about service guarantee effects in the field. More specifically, we seek to understand when, where and how much the received theoretical explanations for service guarantees order the data given the endemic presence of subject and treatment heterogeneity. To accomplish these goals, we develop analysis tools for a multi-site, multi-wave extension of the classic regression discontinuity design (Shadish, Cook and Campbell, 2002). We apply these techniques to data from a mid-priced hotel chain that implemented an SG program to answer the following questions:

1. How do observed and unobserved characteristics of hotels impact the success of the SG program at each site?
2. Which hotel sites should continue or discontinue their SG program?
3. Which hotel sites should be rewarded for their SG implementation?
4. Which remaining hotel sites should be targeted for the SG program implementation?

*Contributions of the paper*

We add to the extant knowledge about service guarantees in the field. Our analysis shows that the effect of the SG program on service evaluation scores varies significantly across sites; 28 hotels displayed significant net gains, 11 hotels displayed significant net declines, and 43 hotels displayed no significant change. Part of this variation is driven by *observed* site characteristics such as pre-existing levels of service quality, and a larger fraction of single-purpose trip guests[1] (both of which prompted gains), but *unobserved* characteristics also play a big role. These patterns are consistent with the theoretical idea of SGs as signals of quality, but not with the view of SGs as incentive devices that motivate employees to deliver superior quality service.

Based on our empirical Bayesian estimates of program effect, we devise policies that respond to the observed and unobserved characteristics of different sites. For instance, applying our program continuation decision rule to the different sites shows that only 13 of the 43 hotels with an insignificant program effect should terminate the program. The remaining 30 hotels should continue with the program *despite the lack of success to date* given their better than even odds of future program success.

We also devise rewards for exceeding expected program impact, and pinpointed 42 hotels that would be rewarded under such a policy. Crucially, our benchmarks incorporate observed and

[1] Business or pleasure trips are single-purpose visits, while multi-purpose trips combine both aspects. The latter visitors are more difficult to serve effectively as they have more varied needs.

unobserved characteristics of each site. Thus, 11 of these rewarded hotels actually exhibited a significantly negative program effect, but are rewarded nevertheless because they performed better than expected for their own location. Likewise, not all hotels with positive program effects would have realized a reward. Indeed, 12 of the 28 hotels with a significantly positive program effect would have gone unrewarded for not exceeding expectations. We stress that these expectations are inferred from the outcome data, and are not obtained from self-reports.

Finally, we identify priority sites for further rollout of the program. Of the 70 hotels that had not yet implemented the SG program, after controlling for each individual hotel's characteristics and history, we are able to pinpoint 17 sites that are very unlikely (<10% odds) to achieve a positive program effect, while 9 other sites are very likely (>90% odds) to do so. Clearly, these latter hotels should be the chain's priority rollout targets.

The remainder of the paper is organized as follows. We first review the relevant literature. Then, we describe our research context. Next, we discuss the methodology employed in our research context, followed by the results. The paper concludes with managerial implications and general discussion.

## LITERATURE REVIEW

A service guarantee (SG) is a set of two promises—a commitment by the firm to make good on a promised level of service and a commitment to compensate the customer when the first promise is not met. The extant work features three mechanisms underlying service guarantees; signaling, risk reduction, and incentives.

Harvey (1998) shows that a SG conveys credible information to customers about the hidden attributes of a service offering. In effect, it is a signal of existing quality. In a closely related argument, Berry & Yadav (1996) hold that a service guarantee reduces customer risk

perception by telling them what would happen when things go wrong. Besides these informational mechanisms, Hays and Hill (2001) hold that that a SG raises employees' motivation, which carries over into actual improvements in delivered service quality. Notice the pragmatic differences to firms. The signaling and risk arguments for SGs make them useful primarily to firms with existing good service quality, whereas the incentive argument makes them useful to all firms as a tool to be used to realize higher levels of service quality levels.[2]

These theoretical insights provide a sound basis for designing SG programs in the field. In addition, the exhortations of industry observers regarding the power of SGs (e.g., Hart, 1988) have spurred a considerable amount of interest in these programs, particularly in the lodging industry. A number of hotel chains have introduced SG programs, but published work remains scarce, so we are still uncertain about the effectiveness of SG programs (e.g., Evans, Clark and Knutson, 1996).

The only two evaluations that we are aware of (Bolton and Drew, 1991; Simester *et al*., 2000) reach ambivalent conclusions about these programs because of subject and treatment heterogeneity issues. The broader literature on program evaluation (e.g., Heckman *et al*., 1997) also cautions that subject heterogeneity and treatment heterogeneity limit the utility of traditional analyses. We address these issues below.

*Subject Heterogeneity*

SG programs are invariably designed to impact self-selected sets of customers. Although observed differences between these self-selected groups can be readily controlled for, the "composition" effect problem arising from unobserved heterogeneity across treated individuals is much more difficult problem. Consider Bolton and Drew's (1991) evaluation of a field

[2] The incentive argument does not imply that SG programs are profitable to all firms. Differences in payout costs matter greatly here.

experiment at a telephone utility, where the firm upgraded its switching equipment at two of its central office sites in order to improve customer service quality. Using a "difference-of-difference"[3] approach, these authors compared consumer perceptions before and after the upgrade from these two experimental sites with comparable data from two other control sites. They concluded that service quality perceptions improved at the individual customer level, but that there was no significant improvement at the office site level.

These contrasting outcomes across levels arise from the unobserved differences in the composition of the customer groups across the four sites. Crucially, this problem would exist even if the experimental and control group sites were to be randomly assigned. Although treatment randomization makes a difference, it is not a panacea as Hutchinson, Kamakura, and Lynch (2000) demonstrate in their paper on bias arising from subject heterogeneity in true experiments. They recommend repeated measures designs to control for unobserved subject heterogeneity, but this runs into several implementation difficulties with field interventions.

First, true field experiments (with randomization) are extremely scarce because of the time and costs involved. For instance, although policy makers consider smaller class sizes to be a core policy issue surrounding the improvement of student performance, Cook (2002) found just six true experiments on this topic in his review covering three decades of work in the field. In our own review of SG programs in the marketing literature and elsewhere, we found no true field experiments at all.

Second, repeated measures designs as advocated by Hutchinson *et al*. (2000) are very difficult to mount in the field because of subject attrition. Some individuals will inevitably drop

[3] A difference of difference estimator compares the before-after increase in the experimental site with the corresponding increase in the control site.

out of repeated treatment occasions notwithstanding the efforts of the experimenter.[4] At best, one is able to implement multi-site designs with non-identical groups of subjects observed over time at each site. As such, the conventional analysis of repeated measures designs that permit us to control for subject heterogeneity needs to be developed further to accommodate quasi-experimental analogs to true repeated measures designs. Our challenge is to develop models and tools for such designs.

*Treatment Heterogeneity*

Unlike carefully controlled laboratory settings designed to minimize treatment heterogeneity, field interventions implement treatments at different sites in different situations at different times. This issue is evident in the Simester *et al*. (2000) evaluation of a service quality intervention. Using a difference-of-difference approach, these authors compared customer satisfaction improvement in test cities against control cities in the US and Spain. They discovered significant program effects in the US cities, but not in the Spanish cities. For their two-wave design, they developed a rather ingenious methodology using non-equivalent dependent variables to control for unobserved heterogeneity across individual respondents. Thus, they were able to conclude that the different results for the different cities were not due to self-selection of individuals. Their pioneering effort points to the significance of accounting for treatment heterogeneity.

*Heterogeneous Causal Effects*

Traditionally, the discovery of different treatment effects across sites has been framed as a matter of generalizability (external validity). Thus, for instance, much of the literature on field

---

[4] This drop-out problem is so prevalent that field program evaluation outcomes often report the "intention to treat effect (ITE)," which includes the drop-outs, instead of the conventional "effect of treatment on the treated (TET)" reported in laboratory analysis. See Shadish, Cook and Campbell (2002) for a full discussion.

versus laboratory experiments pivots on the relative merits of external versus internal validity. Since internal validity is the sine qua non of theory testing work, it is not surprising that randomized laboratory experiments emerge as the gold standard. However, recent work on the philosophy and statistical analysis of causal effects (e.g., Rubin, 1990; Heckman *et al*., 1997) offers a fundamentally different view of treatment effect differences.

In this contemporary view, there are two alternative assumptions about casual effects. In the first instance, a constant causal effect works "… equally for everyone but for random error …" (Hutchinson *et al*., 2000), and our best estimate of this effect is the average outcome difference between subjects who are randomly assigned to a treatment condition versus those who are assigned to a control condition. Thus, between-units designs with random assignment yield the best information about constant causal effects.

In the second view, causal effects are heterogeneous for a variety of reasons, particularly because of unobserved differences. For instance, a drug may have different effects across patients because of (unobserved) genetic differences. The numerous unobserved differences across organizational sites make the heterogeneous causal impact assumption more tenable for field interventions. Crucially, under the heterogeneity assumption, the previous estimator provides very little information about important policy questions; i.e., between-units designs with random assignment are no longer the gold standard.

More specifically, Heckman *et al*. (1997) show that the constant causal effect model cannot address issues such as a) the fraction of people who are made better or worse off by an intervention, b) the identification of sites where an intervention should be continued or discontinued, and c) the identification of promising sites for future implementation of the intervention. They argue that within-units designs with repeated observations are preferable to

between-units designs, because one can develop posterior (Bayesian) estimates of heterogeneous treatment effects for each unit while controlling for unobserved effects. However, off-the-shelf analysis procedures are not yet available for such designs. We apply their general ideas to the particular design used in our SG program implementation.

## RESEARCH CONTEXT

We study a SG program implemented by a mid-priced chain of franchised hotels in North America. The chain offered a service guarantee stated as follows: *Our goal at [hotel name] is 100% guest satisfaction. If you are not satisfied with something, please let us know and we'll make it right or you won't pay.* The program was promoted with hotel signage in lobbies and tent fold cards placed in guest rooms, but no media advertising was used to support the program. The chain offered no formal incentive or reward for program participation, but did reimburse hotels for any program related payouts during the first year[5]. All of the individual hotels were required to participate in a comprehensive training program, complete with training manuals and videotapes, prior to implementation. Invocations of the guarantee were tracked by the chain to provide information to the individual hotels on the reasons for failure and to determine the financial impact of the guarantee program.

Our observation period ranged from January 1998 (the start of the program) to April 1999 and implementation decisions and dates varied across hotels during this period. Of the 188 hotels in the chain, 118 hotels implemented the program during these 16 months.

The hotel organization commissioned a third-party marketing research firm to collect data via telephone surveys of a random sampling of hotel guests. The market research firm collected the data at planned intervals. The questionnaire included customer service evaluation

[5] We lack data on these expenditures, so we are unable to evaluate profitability outcomes.

items and background information, but no personal identifiers. The length of time between surveys as well as the realized sample sizes varied considerably across hotels and time. In total, our data included 85,321 observations from 178 hotels across 16 survey administration dates.[6]

*Suitability of Multi-Level Regression Discontinuity Design*

Figure 1 shows the distribution of program implementation dates. The majority occurred between September 1998 and February 1999. The number of time points for a hotel is the number of survey dates at that site. These observations constitute a regression discontinuity design at each hotel site albeit with different implementation dates, samples and time points across hotels. An important feature is that the program was implemented at each hotel site, and not at the individual guest level. Thus, we have a multi-level regression discontinuity design for these data.

In order to apply a discontinuity design, for each hotel, a sufficient number of survey time points must exist prior to and after the implementation date. This is important because a regression discontinuity design relies on pre-treatment observations to control for unobserved changes. To this end, we selected 81 hotels for analysis, and Figure 2 shows the distribution of the number of their time points. Of these sties, we have 12 or more time points for 72 sites, and as many as 16 time points for 59 sites. Furthermore, we have sufficient time points on both sides of the program date for these sites; 61 of which had at least three observations points before as well as after the implementation. Figure 3 plots the ratio of time points before the program date to the total time points for each hotel. The mean ratio is 0.42, with 60 hotels having ratios between 0.25 and 0.75. Overall, our sample contains fairly balanced numbers of time points

[6] Each of our 85,321 anonymous respondents is considered as a separate guest. The median guest stayed only two times at this chain of 188 hotels over the last 12 months, so repeated surveys from the same guest are a very small (albeit unknown) number in our database.

before and after implementation and appears appropriate for a discontinuity design. Below, we develop the analysis methodology for this multi-site regression discontinuity research design.

## ANALYSIS

Our data possess a three-level nested structure. Each survey questionnaire can be represented by a triple (*j*,*t*,*i*), where *j* = 1, 2,..., 81 hotels, *t* = 1, 2,..., 16 calendar months, and *i*=1, 2,…, 85,321 individual customers. At the lowest level of nesting, individual customers who stay in the same hotel at the same time share common disturbances at both the hotel level and time level. Thus, their evaluation scores correlate more closely than do scores from customers who stay in the same hotel but at different times. In other words, customers are nested within each hotel-time pair (*j*,*t*). At the next level, surveys from the same hotel over time share a common hotel level disturbance. Thus, their evaluation scores correlate more closely than do scores from surveys at different hotels. Therefore, time points (*t*) are nested within hotels (*j*).

We describe our model using four linked equations, although they can be substituted into a single equation:

$$CSE_{jti} = \pi_{0jt} + \pi_1 x_{1jti} + ... + \pi_n x_{njti} + \pi_{n+1} y_1 + ... + \pi_{n+t} y_{t-1} + e_{jti} \quad (1)$$

$$\pi_{0jt} = \beta_{00j} + \beta_{01j} SG_{jt} + r_{0jt} \quad (2)$$

$$\beta_{00j} = \gamma_{000} + u_{00j} \quad (3)$$

$$\beta_{01j} = \gamma_{010} + \gamma_{011} z_{1j} + ... + \gamma_{01n} z_{nj} + u_{01j} \quad (4)$$

where,

$CSE_{jti}$ is the service evaluation score from hotel *j* at time point *t* from customer *i*.

$x_{1jti} \cdots x_{njti}$ are *n* observed characteristics of customer *i* at hotel *j* at time point *t*.

$y_1 \cdots y_{t-1}$ are $t-1$ indicator variables for the *t* time points.

$e_{jti} \sim N\left(0, \sigma^2\right)$ is the hotel, time point, customer-specific error term.

$SG_{jt}$ is an indicator variable of the SG program status for hotel $j$ at time point $t$.

$r_{0jt} \sim N\left(0, \rho^2\right)$ is the hotel-time specific error term.

$z_{1j} \cdots z_{mj}$ are $m$ observed characteristics of hotel $j$.

$\begin{bmatrix} u_{00j} \\ u_{01j} \end{bmatrix} \sim \mathbf{N}(\mathbf{0}, \mathbf{T})$ is the matrix of the hotel-specific error terms. Denote the variances of $u_{00j}$ and $u_{01j}$ as $\tau_{00}^2$ and $\tau_{01}^2$ respectively.

Equation (1) models the customer service evaluation score ($CSE_{jti}$) for customer $i$ at time $t$ in hotel $j$ as a function of the hotel-time effect (the random intercept term, $\pi_{0jt}$), the **X** vector of $n$ characteristics of that customer, the **Y** vector of $t-1$ indicator variables of time points, and a customer level error term.[7] Equation (2) models the hotel-time effect (the random intercept from the previous equation) as a function of the hotel effect (the random intercept, $\beta_{00j}$), an indicator variable capturing program status ($SG_{jt}$), and a hotel-time level error term. Equation (3) models the hotel effect (the random intercept from the previous equation) as a function of its grand mean ($\gamma_{000}$) and a hotel level error term. Equation (4) models the hotel-specific SG program effect (the random coefficient of $SG_{jt}$ from equation 2) as a function of its grand mean ($\gamma_{010}$), the **Z** vector of $m$ characteristics of the hotel, and a hotel level error term.

The parameters to be estimated are the fixed coefficients $\{\pi_1 .... \pi_{n+t}, \gamma_{000}, \gamma_{010}, \gamma_{011}, ... \gamma_{01m}\}$, the random coefficients $\{\pi_{0jt}, \beta_{00j}, \beta_{01j}\}$, and the parameters of the distribution of the random effects $\{\sigma, \rho, T\}$. Of particular interest is the $\gamma_{010}$ term, which represents the grand mean of the

[7] Serial correlation is not modeled, but we think this is not a large issue here as random samples are taken monthly from each hotel, and the inter-purchase interval between hotel stays is quite large.

heterogeneous SG program causal impact, and the $\beta_{01j}$ terms, each of which represents the SG causal effect at hotel *j*. The EM algorithm (Dempster *et al*., 1981) is used to estimate (a) the fixed coefficients and the variances of the prior distributions $\{\sigma, \rho, T\}$ via Full Information Maximum Likelihood, and (b) the random coefficients via an empirical Bayes method. In the latter instance, it is the posterior distributions of these random effects that are computed. We refer the reader to Bryk and Raudenbush (1993) for computational details.[8]

*Identification*

The program effect is identified because we observe CSE on multiple occasions before and after the implementation at each site. We control for site-specific causal factors via the observed site characteristics and the unobserved site-occasion error term, so we can pinpoint the causal effect at that site arising from the SG program by comparing the before and after scores. Similarly, the identification of the effects of the moderating factors at the customer and hotel levels derives from comparing across hotels and customers respectively. As above, the inclusion of the hotel-level and customer-level error terms that control for unobserved difference

*Outcome Measure*

Our outcome variable is the Customer Service Evaluation (CSE) scale, which is constructed from the following survey questions:

Q1: How likely would you be to stay at this (hotel chain name) again? (1 to 5 scale)

Q2: How likely would you be to recommend this specific (hotel chain name) to a friend? (1-10 scale)

Q3: Value per price paid represented by this (hotel chain name) stay. (1-10 scale)

---

[8] The prior distribution of $\beta_{01j}$, which is our random causal impact, can be written as
$\beta_{01j} \sim N\left(\gamma_{010} + \gamma_{011} z_{1j} + \gamma_{012} z_{2j} + \ldots + \gamma_{01n} z_{nj}, \tau_{01}^2\right)$

Q4: How would you rate your overall satisfaction? (1-10 scale)

We summed the items to create the CSE scale.[9] The psychometric quality of our summed scale is assessed via factor analysis. The scree plot in Figure 4 shows that one factor suffices to explain the variation in these data. Table 1 shows that each of the items loads strongly on the single factor.

*Observed Customer Characteristics*

Questionnaire items about the purpose of the trip and previous nights stayed at this brand were used to construct the following measures:

$BUS_{jti}$: Indicator variable set to 1 if the purpose of the trip was business only and zero otherwise.

$VAC_{jti}$: Indicator variable set to 1 if the purpose of the trip was vacation only and zero otherwise.

$BRANDLOYAL_{jti}$: Share of nights stayed at this hotel brand over the past 12 months. This was constructed from two questions asking the customer (a) the total number of nights stayed at any hotel, and (b) the number of nights stayed at this hotel brand over the past 12 months.

*Time Effects*

Archival data from the hotel chain were used to construct the following 15 dummy variables to capture unobserved effects occurring over time.

$Month_t$: 15 indicator variables to indicate the 16 months from January 1998 to March 1999. The base month is set as April 1999 (Month 16). Each variable was set to 1 if the survey observation occurred in that month and zero otherwise.

*SG Program Status*

Archival data from the hotel chain was used to construct an indicator variable ($SG_{jt}$), which was set to 1 if the SG was in effect in hotel $j$ at time $t$ and zero otherwise.

---

[9] Q1 was rescaled to conform to the response scale of the other items.

*Observed Hotel Characteristics*

Our hotel characteristics measures are averaged responses from observations of individual guests from surveys conducted at time points prior to the SG implementation date for that site, which prevents confounding of intervention effects with measured characteristics.

$CSE\_HPRE_j$: CSE score for hotel $j$ averaged over guests and times points prior to its SG implementation.

$BUS\_HPRE_j$: The proportion of customers at hotel $j$ on business trips averaged over guests and time points prior to its SG implementation.

$VAC\_HPRE_j$: The proportion of customers at hotel $j$ on vacation trips averaged over guests and time points prior to its SG implementation.

$BRANDLOYAL\_HPRE_j$: The share of nights the customer stayed at this brand averaged over guests and time points at hotel $j$ prior to its SG implementation.

*Sample Characteristics*

Table 2 reports the descriptive statistics of our sample. Some of the average values of the customer characteristics measures are worth noting. A large percentage of customers (87%) stayed at the hotels on a single purpose trip, with 47% on vacation trips and 40% on business trips. The remaining 13% of the customers were on dual purpose. We speculate that customers on single purpose trips are easier to serve than are customers on dual purpose trips, which implies positive coefficients for $BUS_{jti}$ and $VAC_{jti}$ in equation 1. In terms of the share of nights stayed, the data show a fairly loyal customer base, averaging 0.35. That is, on average, a current customer spent approximately one third of her/his travel nights at this hotel brand in the past year. For the lodging industry, this is a respectable level of loyalty. Intuitively, one would expect more loyal customers to rate the service higher, which implies a positive coefficient for

$BRANDLOYAL_{jti}$ in equation 1. Turning to the critical hotel-time measure of program status, the mean of $SG_{jt}$ is 0.42, which shows that that we have a fairly balanced set of pre- and post-program data.

Some of the observed hotel characteristics are also noteworthy. First, this hotel chain enjoyed a fairly strong customer base prior to the program. The chain level average of $CSE\_HPRE_j$ is 30.2, which is around 7.5 on a 10-point scale. To put this into perspective, note that the response anchors for the original 1-10 scale were as follows: 5-7 is "fair" and 8-10 is "excellent." The scores range from 24.81 to 34.20 across hotels, which is approximately 6 to 8 on a 1-10 scale. Hence, even the worst hotel in the survey had "fair" scores before program implementation. The customer loyalty measure prior to program implementation reveals a similar pattern. $BRANDLOYAL\_HPRE_j$ ranges from 0.282 to 0.458, which implies that in the worst case a customer spent nearly 30% of travel nights in this hotel chain in the past year.[10]

There is considerable heterogeneity across the hotel sites, particularly in terms of the purpose of trip. For example, one hotel has only 7.5% of its guests on business trips, while another one has 71.7%. Similarly, one hotel has only 15.5% of its guests on vacation trips, while another hotel has 78.4%. This leads us to speculate that the service guarantee program might well have very different effects across hotels.

*Endogeneity*

In our data, the $brandloyal_{jti}$ measures the "share-of-nights" of the chain in a customer's total hotel stays in hotel in the 12 months *prior* to the date of the survey. One concern is that this measure could be endogenous, i.e., for any individual customer, the endogeneity problem arises when his report on this variable reflects stays at the chain that resulted from the SG program.

[10] Parenthetically, this is much higher than the share of requirements measures of loyalty typically found in consumer packaged goods.

An econometric response would embed a model of the customer into the estimation. However, this requires us to make a number of additional assumptions as well as to obtain data on these presumed drivers of customer decisions. In our quasi-experimental approach, we appeal to features of our design to rule out endogeneity in much the same way as a true experiment. There are several pieces of evidence below that converge on this point.

First, this hotel chain did not advertise the SG, and used only in-hotel signage and tent-fold cards for promotion; it is very likely that customers become aware of the SG program only when they stay at a post-implementation hotel. Given this observation, endogeneity issues are confined to those observations from customers who stayed at least 3 times at a given site, and who completed a survey on or after their second post-SG stay. Note that 75% of hotels are observed for at most 8 months after SG implementation. Let us use this as a baseline. If a customer were to have stayed at the same hotel at least three times after its SG start date, the implied average frequency of stays in the same hotel is $12 \div (8/3) = 4.5$ per year. Inspecting our data, we found that only 5% of the surveyed customers fell into such a category. In addition, for each customer survey from a given hotel after its SG start date, let $q$ denote the frequency of stay in the same hotel in the last 12 months, and $t$ denote the time interval (years) between the survey date and the SG start date. For $(t - 2/q) > 0$, it is possible that this customer came to this hotel more than twice after SG implementation, and the surveyed stay may well have been the third or more post-SG stay at that hotel. It turns out that we have only 638 customer surveys that belong to this category (3.59% of observations) of potentially "contaminated" observations. We reran our model after dropping these observations, and obtained the same results.

Second, it is possible that a customer stayed at one hotel post-SG, and then chose to stay at another site in the same chain because of his favorable impression of the program.

Endogeneity would be a concern for these customers prior to their third post-SG stay at the same site. We think this chain of events is rare given that a) the chain doesn't advertise the SG program, and b) each hotel implemented its SG programs at different time points in a manner that is not transparent or predictable to customers.

Third, we considered the problem as instrumental variable problem (in reverse). Suppose that our *brandloyal* measure were endogenous; one should be able to verify this by regressing *brandloyal* against SG and other known exogenous variables. This regression yielded an insignificant negative SG coefficient.

In sum, endogeneity does not appear to be a problem in our quasi-experiment; of course, one cannot rule it out completely without a true experiment.

## RESULTS

As a first-cut analysis, each hotel is analyzed as an observation unit from a quasi-experimental design consisting of a single within-subject factor. This analysis does not control for either subject or treatment heterogeneity. A paired *t*-test on the CSE observations for this design shows that the SG program *lowered* customer evaluations ( $CES_{before} - CES_{after} = 0.42$ ; $p<0.01$. Parenthetically, the chain used precisely this before-after design to evaluate the program among its early adopter sites.

Next, we analyzed these data with a classic regression discontinuity design (e.g., Shadish, Cook, and Campbell, 2002). The average CSE score at time point *t* at hotel *j* is modeled as a function of customer group characteristics, program status and time effects as follows:

$$CSE_{jt} = \alpha_0 + \alpha_1 BUS_{jt} + \alpha_2 VAC_{jt} + \alpha_3 BRANDLOYAL_{jt} + \alpha_4 TIME_{jt} + \alpha_5 SG_{jt} + r_{jt} \qquad (5)$$

where $CSE_{jt}$, $BUS_{jt}$, $VAC_{jt}$, and $BRANDLOYAL_{jt}$ are customer characteristics measures aggregated to hotel *j* at time *t*. $TIME_{jt}$ is the number of days elapsed from the program start date

to a survey at hotel *j* at time *t*. It is a negative number for *t* prior to implementation and positive after implementation, and controls for unobserved time effects. Finally, the *SG* measure captures program status, and its coefficient estimates the constant causal effect after controlling for these other effects.

Contrary to the result above, this analysis (Table 3) yields a positive, albeit insignificant SG program effect. Both analyses ignore potential treatment and subject heterogeneity, which as we have seen can yield misleading estimates. It highlights the need to address these sources of heterogeneity we include in our model (Equations 1-4). Below, we present the results from this specification. We discuss the variance components first and then discuss results starting from the customer level, and then work up the levels of the nested structure. Finally, we discuss the SG program effects of central interest to us.

*Variance Components*

One of the benefits of our model is that it estimates the variance in the outcome variable that is located at each level of our nested structure. Table 4 shows that the standard deviation of our CSE measure between individual customers ($\sigma$=7.853) is approximately eight times as large as the corresponding dispersion across hotel-time occasions ($\rho$ = 0.962) or across hotels ( $\tau_{00}$ = 1.977). Substantively, this highlights the multiple loci of variation in customer service evaluations. Each individual's experience and background, the context of the particular visit, and the enduring aspects of that hotel all influence his evaluation score[11]. The sources of these influences are seen more clearly by examining the relevant coefficients.

[11] Managerially, the large variance located at the individual customer level speaks to the need to market the program directly to customers in addition to site level efforts.

*Customer Effects*

As we suspected, there are differences in CSE responses across travelers on different trip types. Single purpose travelers' service evaluations are higher than the ratings from dual purpose travelers, as evidenced by the positive significant coefficients for our two indicator variables ($BUS_{jti}$ and $VAC_{jti}$) measuring differences against the base (dual purpose trip) category. Notwithstanding these statistical differences, these differences across trip purposes are quite small in terms of absolute magnitude. On the 4-40 CSE scale, the largest difference among different trip purposes is between personal trips and dual-purpose trips ($\pi_2 = 0.47$, $p < 0.05$; $\pi_3 = 4.035$, $p < 0.05$). Projected to the 1-10 scale, this is only a score difference of 0.1.

The $BRANDLOYAL_{jti}$ variable shows a positive significant effect on CSE evaluations. Again, this is intuitively plausible as individuals who tend to stay disproportionately at this chain are more likely to have been satisfied with their experience. This effect is also larger in magnitude than the trip purpose effect. To put this effect into perspective, denote a customer who stayed exclusively at this hotel brand during the past year as a "Fully Loyal Customer." Likewise, denote a customer with no previous stays at this hotel brand in the past year as a "New Customer." From Table 4, we compute the CSE score difference between a fully loyal customer and a new customer to be about 1 on a 1-10 scale, which is quite large. We hasten to add that this loyalty effect is not the SG program effect. However, it hints at the issue that a hotel with more loyal customers starts from a higher base of evaluation scores.

*Time Effects*

Six of the fifteen coefficients for the $Month_t$ indicator variables are significantly different from the baseline (Month 16, April 1999). Figure 5 plots the effects of unobserved monthly events. Observe that CSE scores declined in the first quarter of 1998 and then became relatively

stable until October of that year. From October to December, scores climbed back to the level of the start of the year. There is a very large drop in January 1999, followed by a recovery, suggesting that some large unobserved event influenced customer evaluations at the end of 1998. These effects underscore the importance of controlling for unobserved changes through repeated observations of the same unit.[12]

*Hotel Characteristics*

In Equation 3, the hotel effect consists of a fixed part (grand mean) and a random part. In Table 4, we see that the grand mean ($\gamma_{000}$=28, $p$<0.05) is quite large in magnitude. It projects to a 7 on the 1-10 scale, which reinforces the earlier assertion that this hotel chain performs quite well on average. However, there are significant differences across these hotels as evidenced by the significant estimate of the standard deviation of $u_{00}$ ($\tau_{00}$=1.977, $p$<0.05).

*SG Program Effects*

In Equation 4, the SG program effect consists of a fixed part (the grand mean and a linear combination of the observed hotel characteristics), and a random part. Our grand mean estimate in Table 4 is not significantly different from zero ($\gamma_{010}$=0.138, p> .10); however, this does not imply that the program effect is uniformly insignificant across hotels because the observed hotel characteristics have significant effects.

Hotels with better CSE scores prior to program implementation benefit more from the SG program ($\gamma_{011}$= 0.162, $p$<0.05). This result comports with the mechanism from theoretical work that a service guarantee acts a signal conveyed by the hotel to its guests (e.g., Boulding and Kirmani, 1993). Whether such a signal works or not depends highly on the credibility of the

[12] There was no substantive explanation available to account for this one-off event. It also supports our choice of a fixed effects dummy variable specification for these unobserved effects over a random effect specification.

source. In our context, a hotel with a better service history (i.e., a higher CSE_HPRE score), is more credible than a hotel with a poorer service history. At the former hotel, customers believe the SG claim and increase their CSE evaluations of their visit. At the latter hotel, customers discount the SG claim and lower their CSE evaluations.

This result does not comport with the mechanism posited by Hays and Hill (2001) to the effect that service guarantees motivate the delivery of better service. In our results, the SG program does not work to salvage poorly functioning hotels, but does serve to signal good service hotels. However, this does not mean that the latter is not a theoretically relevant mechanism; rather, the signaling mechanism appears to be dominant in this context.

Table 4 shows that hotels with larger fractions of single purpose travelers prior to program implementation benefit more from the SG program ($\gamma_{012} = 7.706$, $p<0.01$; $\gamma_{013} = 8.194$, $p<0.01$ for business and vacation trip fractions, respectively). This result comports with the signaling mechanism because easier-to-serve guests are less likely to encounter the promise contained within the signal being subsequently disconfirmed by a negative service encounter.

Finally, the average share-of-stays of a hotel's customer base prior to program implementation did not have a significant influence on the causal impact of the SG program ($\gamma_{014} = 1.302$, $p>0.10$). In other words, the loyalty of a hotel's customer base has no effect on that hotel's SG program impact. Recall, however, that we did find a significant positive effect of an individual customer's loyalty on his/her CSE score. Put differently, we find a *compositional* effect, but no *contextual* effect of brand loyalty.[13] The intuition behind these effects of loyalty is

[13] Compositional effects are the aggregation of individual level causal effects (such as perceptions, attitudes, etc.) that operate on an individual, while contextual effects are the impact of the group context on the same individual (e.g., size of group, etc.).

as follows. Customers who often stay at this hotel chain are unmoved by service guarantee claims of individual hotel sites, but they are influenced by their individual experiences.

*Hotel-Specific Program Effects*

The central issue of our inquiry is the heterogeneous causal impact of the SG program. Thus far, we have shown that the grand mean of the causal impact across hotels is not significant; however, both prior service quality history and trip purpose characteristics of the customer base of a hotel significantly impact the causal effect of the SG program at that hotel. In addition to these fixed components of program effect, we need to add the random part (realization of $\tau_{01}$) to arrive at the net effect of the SG program at a given hotel. Using Bayesian procedures, we get the posterior estimate, $\overline{\beta}_{01j}$, that describes the net effect of the SG program at hotel *j*.

Due to space constraints, we do not report the 81 individual estimates, but Table 5 shows our posterior estimates varying from –0.99 to 1.166. We plot these estimates in Figure 6. There is a right-skewed distribution, with most hotels showing program effects between 0 and 0.6.

Given the distribution of these estimates, $N\left(\overline{\beta}_{01j}, \lambda^2\right)$, we assess the statistical significance at each hotel by computing $\overline{\beta}_{01j} / \lambda$ .When this ratio exceeds 1.28, there is a 90% chance that $\beta_{01j}$ is greater than zero at that site. Likewise, when $\left(\overline{\beta}_{01j} / \lambda\right) < -1.28$, there is a 90% chance that $\beta_{01j}$ *is* less than zero.

In Figure 7, we see that of the 58 hotels with a positive posterior effect, 28 of them exceed a 90% threshold for a positive causal impact, while the remaining 30 hotels exceed a 50% threshold. On the other hand, 23 hotels have a negative posterior effect, of which 11 hotels exceed a 90% threshold, while the remaining 12 hotels exceed a 50% threshold level.

To sum up, we find large variability in SG program outcomes across the individual hotels. Some part of this variation is predictable given a hotel's prior observed level of quality of service and its prior observed fraction of single-purpose trips, while the remainder arises from unobserved differences across hotels. Thus, policies that apply to all hotels, or even policies that adjust for a hotel's observed profile can be improved upon by using hotel-specific posterior estimates to formulate site-specific policies that account for the unobserved differences.

## FORMULATING MANAGERIAL DECISION RULES

### *Stop/Continue Rule*

An obvious question facing the chain is whether to continue with the SG program at each implemented site. We develop a rule as follows. For each hotel with an SG program in place, we calculate the probability of a positive causal effect in the manner described above. Next, we impose a probability threshold. Thus, with a 90% threshold rule, one would conclude that the 28 hotels in Area IV of Figure 7 should continue with the SG program, while the 11 hotels in Area I of Figure 7 should discontinue the program. For the 42 hotels in the middle (Areas II and III), the decision rests on the chosen threshold. For example, if the chain decides that a greater than 50% chance of program success is acceptable, the 30 hotels in Area III should continue with the program, while the 12 hotels in Area II should stop. This richer decision rule is based on more information than an invariant stop/continue policy implied by a constant casual effect model (such as might be derived from our initial analyses of the program).

### *Reward Policies*

Recall that the chain depended on each hotel to implement the program at its own site. As such, incentives for performance are an important managerial tool, but geographic dispersion makes it prohibitively costly for the chain to observe site managers' efforts directly. Equally

important, independent hotel owners are often unwilling to allow headquarters to use subjectively determined bonuses. In sum rewards based on direct observations of effort, and/or subjective bonuses are not feasible, so it is useful to consider rewards based on observed, albeit noisy outcomes such as CSE. We know from theoretical models of asymmetric information that noisy measures are useful inputs into designing rewards (e.g., Holmstrom and Milgrom, 1991). Our key challenge is to develop a policy that is both fair and easily communicated to participants. Intuitively, policies should compare a hotel's outcome against its own history as well as against other similarly situated hotels. As such, consider policies based on observed as well as unobserved differences. Such policies are more inclusive than a common policy (which ignores all differences related to CSE outcomes) or policies that adjust only for observed characteristics (which ignores all unobserved differences related to CSE outcomes).

*Exceeding Expectations Policy 1*: Consider a simple policy that rewards all hotels with a significant positive posterior estimate $\overline{\beta}_{01j}$; this is our best estimate of SG program effect, and it incorporates both the observed and unobserved characteristics of each site. In Figure 7, there are 28 such hotels. Unfortunately, this policy would still discourage managers of poor hotels because positive results are harder for them to achieve, while it over-rewards managers of good hotels whose outcomes are so much easier to achieve.

Consider the baseline for each site as the program effect adjusted for observed characteristics. We can compute this as $\gamma_{010} + \gamma_{011} z_{1j} + \gamma_{012} z_{2j} + ... + \gamma_{01m} z_{mj}$. Denote this as our expectation of program impact ($\tilde{\beta}_{01j}$). Next, considering rewarding all hotels with a positive value of $\overline{\beta}_{01j} - \tilde{\beta}_{01j}$. In Figure 8, there are 42 hotels (Areas I, II and III) with positive values of this difference

Figure 8 reveals a subtle aspect of this policy. By plotting $\overline{\beta}_{01j} - \tilde{\beta}_{01j}$ against $\left(\overline{\beta}_{01j} / \lambda\right)$, we see 11 hotels (Areas I and IV) where our policy decision above was to discontinue the SG program because they all exhibited a significantly negative program impact (below the 90% threshold level). However, 6 of these hotels (Area I) exceeded our expectation, and therefore would be rewarded *despite the decision to discontinue*. On the other hand, the program was to be continued at the 28 hotels (Areas III and VI) with a significantly positive program impact (above the 90% threshold level), but that 12 of these hotels (area VI) performed below expectation, and thus would not be rewarded *despite the continuation decision*. This policy serves to remove some of the organizational stigma attached to failed initiatives, which tends to create risk-averse managers.

*Exceeding Expectations Policy II:* One limitation of the reward policy described above is that it ignores the uncertainty (posterior variance) in our program effect (posterior mean) at each hotel. To see this, observe that the prior and posterior distributions of SG program effect are as follows:

Prior Distribution: $\beta_{01j} \sim N\left(\gamma_{010} + \gamma_{011} z_{1j} + \gamma_{012} z_{2j} + ... + \gamma_{01m} z_{mj}, \tau_{01}^2\right)$

Posterior Distribution: $\beta_{01j} \sim N\left(\overline{\beta}_{01j}, \lambda^2\right)$

Denote $P_0$ and $P_1$ as the probabilities of program effects greater than zero given our estimates of the prior and posterior distribution, respectively. Since the posterior variance is always smaller than prior variance, it may well be the case that the posterior mean is positive and smaller than the prior mean, but that $P_0 < P_1$. Likewise, the posterior mean might be negative and greater than the prior mean, but that $P_0 > P_1$.

We consider a policy that rewards sites with a positive difference in these probabilities. The probability calculated from the prior distribution is the prior belief about the success of a SG program at a given hotel, while the probability calculated from the posterior distribution is our posterior belief about program success. Rewards go to those hotels that show increases in these probability estimates.

Using our model's estimates, we can readily calculate the relevant probabilities. Figure 9 plots the computed probability difference ($P_1 - P_0$) for each hotel. Under this policy, the 46 hotels with positive values of $P_1 - P_0$ (Areas I, II and III) would be rewarded.

Contrasting it with the first policy depicted in Figure 8, we see several shifts. Of the 11 hotels with significantly negative program effects (Areas I and IV in Figure 8), the four hotels from Area 1 now move to Area IV in Figure 9. That is, even though these four hotels would have been rewarded under the previous policy, they will not be rewarded under the revised policy. As for the 42 hotels with insignificant program effects (Areas II and V in Figure 8), two previously rewarded hotels from Area II now move into the unrewarded Area V of Figure 9. Likewise, two previously unrewarded hotels from Area V in Figure 8 move into the rewarded Area II of Figure 9. Finally, of the 28 hotels with positive program effects, eight previously unrewarded hotels (Area VI in Figure 8) now move into the rewarded Area III of Figure 9. Overall, this revised reward policy corrects the downward bias in the previous policy, albeit at the cost of reduced transparency because of these computed probabilities.[14]

---

[14] A question that arises is whether our previous rules for making stop/continue decisions and our reward policies all favor the same hotels. In other words, "what hotel is more likely to get rewarded -- a poor hotel, a good hotel, or does it not make a difference?" To this end, we regressed our posterior-prior probability difference $(P_1 - P_0)$ against the prior mean for program effect for each hotel, e.g., $\left(P_1 - P_0\right)_j = \alpha + \delta\tilde{\beta}_{01j} + \varepsilon_j$. This regression yielded a negative estimate of the slope, δ, which suggests that a poor hotel is *more* likely to achieve the same level of gain in success rate than is a good hotel.

Although these reward policies use information about observed and unobserved effects, and are thus more appealing intuitively, there are two notes of caution. First, it is important to gain legitimacy from managers for these more complex policies, so the simpler policies may be preferred in some instances. Second, the effects of these complex policies need to be established empirically as incentive policies do not always work out in the field.

*Targeting Future Program Sites*

There are 70 hotels in our data that had not yet implemented the SG program. Which of these hotels should the chain target? Plainly, it pays to target the hotels with the highest expected program impact, which in our model is represented as $\beta_{01j} \sim N\left(\tilde{\beta}_{01j}, \tau_{01}^2\right)$. Applying equation 4, we compute this quantity for each of the 70 hotels as follows:

$$\begin{aligned}\tilde{\beta}_{01j} = {} & 0.138 + 0.162\left(CES\_HPRE_j - 30.199\right) + 7.706\left(BUS\_HPRE_j - 0.46\right) \\ & +8.194\left(VAC\_HPRE_j - 0.41\right) + 1.302\left(BRANDLOYAL\_HPRE_j - 0.352\right)\end{aligned} \quad (6)$$

The numbers in the parentheses are the means of the corresponding variables for the 81 hotels that have already implemented the program. Figure 10 plots these quantities against $\left(\tilde{\beta}_{01j} / \tau_{01}\right)$, and we find that there is less than 10% chance of success at 17 hotels (Area I), so the SG program should not be rolled out to these hotels. Conversely, there is a better than 90% chance of success at the 9 hotels in Area IV; thus, they should be targeted for implementation. The hotels in Areas II and III are more difficult to classify. A conservative decision maker might target the 17 hotels in Area III that show a better than even chance of a positive SG program effect, whereas a bolder decision-maker might accept lower odds of program success and target more hotels.

## DISCUSSION

This paper presents results from a service guarantee program implemented at 81 hotels (but not yet implemented at 70 other hotels) of a mid-priced hotel brand. Using our extension of a regression discontinuity design, and a randomly varying causal effect specification, we estimated the program impact at each site while controlling for observed and unobserved characteristics. Armed with these estimates, we address several conceptual and managerial questions that are important to the firm's ability to learn from such an intervention.

First, we unpacked the factors that drive the impact of this SG program. It does significantly better at hotels with better pre-existing service evaluation scores as well as at hotels with easier to serve guests (single-purpose trips). In terms of the mechanisms that have been implicated in theoretically focused laboratory studies, these results suggest that SG works as a signal of good service, but that it does not appear to work as a tool to improve service delivery. We hasten to add that our analysis does not seek to pit one causal story against another, but rather to unpack where, when and how this program works.

Second, we devise managerial polices that are responsive to the observed and unobserved characteristics of each hotel. A stop/continue decision rule based on our posterior estimates of program effect identifies 11 hotels that should discontinue the program as well as 28 hotels that should continue with the SG program. The decisions at the remaining 42 hotel sites depend on the threshold of success required by the decision maker.

We also devise policies that reward hotels for exceeding expectations at a site. These policies respond to the need to employ observed outcome measures, given the practical difficulties of direct monitoring of effort at far-flung sites, but which are also responsive to the observed and unobserved characteristics of each site. As such, our policies reduce the risk that

owners and managers of "more difficult" hotels might be punished for circumstances that are out of their control. Our policies reduce the stigma of failure and inspire risk-taking because hotels with a negative program impact can still be rewarded provided they exceed expectations. Indeed, we found that even hotels where the programs were to be discontinued because of negative impact can still gain from these reward policies. It should be stressed that we did not use self-reported measures from customers or managers about their expectations but, instead, infer them from our estimates.

Finally, we devise a targeting strategy for program rollout. We apply the model's estimated parameters to the data from the 70 hotels that have not yet implemented the SG program in order to identify the most promising sites. Using our expected program impact estimates, we identify 9 target hotels where the predicted probability of a positive program impact exceeds 90%. Another 17 hotels are expected to show poor program effects (below a 10% probability of success), so they should be discouraged from using the program. However, as Heckman *et al*. (1997) stress regarding similar decisions about social program eligibility, it may nevertheless be beneficial to offer the managers of these latter hotels the option of rolling out the program in order to accommodate the risk-takers among them. This is particularly relevant for the 44 sites with borderline predicted program impact.

*Conclusions*

Despite the call for a more nuanced view of causal impact by marketing scholars (e.g., Hutchinson *et al*., 2000) and the need for more comprehensive models of program evaluation by social policy analysts (e.g., Heckman *et al*., 1997), there has been little progress in the analysis of field quasi-experiments (see Simester *et al*., 2000 for a notable exception) along these lines. Much of the debate about field studies still revolves around the supposed superior external

validity of field studies versus the superior internal validity of laboratory work. We refashion this debate by focusing on the ramifications of constant versus randomly varying causal effects. Our success in improving our conceptual understanding of when, where and why the SG program works, as well as our ability to devise managerial policies that are sensitive to the fundamental problems of site and treatment heterogeneity leave us optimistic about the utility of these newer approaches to address previously unanswered questions.

Some fruitful avenues for future research are also indicated by the current work. One limitation of the current study is the problem of the hidden effort on the part of the hotel manager, as viewed from headquarters. If the effort interacts with hotel specific features, our estimate of program effect could be biased downwards given our dummy variable regressor.[15] Hence, it would be useful to validate our method with measures of SG effort at each hotel. Development of the varying causal effect paradigm appears to be a priority item as it is the building block for our work. Marketing has lagged behind the program evaluation field in this respect, and we can usefully adapt their work to other quasi-experimental designs, particularly commonly used cross-sectional designs where selection issues are more prominent. From a managerial standpoint, the next step is to test the alternative reward policies developed in our paper because they are quite complex, and it is entirely possible that managers who do not understand the manner in which unobserved site-specific differences are controlled in the analysis might well feel resentful as they see sites that are less successful than their own site be rewarded. Communicating the site-specific expectations underlying our reward policies will require field testing and experimentation. Another direction is to deal squarely with our lack of

[15] This was brought to our attention by a reviewer.

sales and profit outcome data, which is a long standing gap in the marketing literature. Hopefully, future work will pick up these challenges.

REFERENCES

Anderson, Eugene W., Claes Fornell, and Donald R. Lehmann (1994), "Customer Satisfaction, Market Share, and Profitability: Findings from Sweden," *Journal of Marketing*, 58 (July), 53-66.

---------- and Mary W. Sullivan (1993), "The Antecedents and Consequences of Customer Satisfaction for Firms," *Marketing Science*, 12 (Spring), 125-43.

Berry, Leonard L. and Manjit S. Yadav (1996), "Capture and Communicate Value in the Pricing of Services," *Sloan Management Review*, 37 (4), 41-51.

Bolton, Ruth N., James H. Drew (1991), "A Longitudinal Analysis of the Impact of Service Changes on Customer Attitudes," *Journal of Marketing*, 55 (January) 1-9.

Boulding, William and Amna Kirmani (1993), "A Consumer-Side Experimental Examination of Signaling Theory," *Journal of Consumer Research*, 20 (1), 111-123.

Bryk, Anthony S., Stephen W. Raudenbush (1992), *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, CA, Sage.

Cook, Thomas D. (2002), "Randomized Experiments in Education: Why are they so Rare?" Working Paper (WP-02-19), *Institute for Policy Research*, Northwestern University.

Evans, Michael R., Dana J. Clark, and Bonnie J. Knutson (1996), "The 100-Percent, Unconditional, Money-back Guarantee," *Cornell Hotel and Restaurant Administration Quarterly*, 37 (6) December, 6-7.

Fornell, Claes (1992), "A National Customer Satisfaction Barometer," *Journal of Marketing*, 56 (January), 6-21.

Hart, Christopher W.L. (1988), "The Power of Unconditional Service Guarantees," *Harvard Business Review*, 66, (4), 54-62.

Harvey, Jean (1998), "Service Quality: A Tutorial," *Journal of Operation Management*, 16, Issue 5 (October) 583-97.

Hauser, John R., Duncan I. Simester, and Birger Wernerfelt (1994), "Customer Satisfaction Incentives," *Marketing Science*, 13 (Fall), 327-50

------- (1996), "Internal Customers and Internal Suppliers," *Journal of Marketing Research*, 33 (3), 268-80.

------- (1997), "Side Payments in Marketing," *Marketing Science*, 16 (3), 246-55.

Hays, Julie M., and Arthur V. Hill (2001), "A Longitudinal Study of the Effect of a Service Guarantee on Service Quality," *Production and Operation Management*, 10 (4), 405-423.

Heckman, James J. (1976), "Shadow Wages, Market Prices, and Labor Supply," *Econometrica*, 46, 403-426

----------, Jeffrey Smith and Nancy Clements (1997), "Making the Most Out Of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64, 487-535.

Holmstrom, Bengt, and Paul Milgrom (1991), "Multi-task Principal Agent Analyses: Incentive Contracts, Asset Ownership and Job Design," *Journal of Law, Economics And Organization*, 7, 25-52.

Hutchinson, J. Wesley, Wagner A. Kamakura and John G. Lynch (2000), "Unobserved Heterogeneity as an Alternative Explanation for 'Reversal' Effects in Behavior Research," *Journal of Consumer Research*, 27 (3), 324-45.

Rubin, Donald B. (1990), "Formal Modes of Statistical Inference for Causal Effects," *Journal of Statistical Planning and Inference,* 25, 279-92.

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

Shadish, William R., Thomas D. Cook, and Donald T. Campbell (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

Simester, Duncan I., John R. Hauser, Birger Wernerfelt and Roland T. Rust (2000), "Implementing Quality Improvement Programs Designed to Enhance Customer Satisfaction: Quasi-Experiments in the United States and Spain," *Journal of Marketing Research*, 37, No. 1 (February) 102-12.

Table 1: Factor analysis of CSE Items

| Variable | Factor Loading |
|---|---|
| Q1 | .747 |
| Q2 | .922 |
| Q3 | .844 |
| Q4 | .877 |

Extraction Method: Principal Axis Factoring. 1 Factor extracted.

Table 2: Descriptive statistics

| Variable | Observations | Minimum | Maximum | Mean | Standard Deviation |
|---|---|---|---|---|---|
| Customer Service Evaluation $CSE_{jti}$ | 49131 | 4 | 40 | 29.94 | 8.54 |
| Customer's Vacation Trip Dummy $VAC_{jti}$ | 46986 | 0 | 1 | .47 | .50 |
| Customer's Business Trip Dummy $BUS_{jti}$ | 46986 | 0 | 1 | .40 | .49 |
| Customer's Share of Stays $BRANDLOYAL_{jti}$ | 47248 | 0 | 1 | .35 | .27 |
| SG Program Status Dummy $SG_{jt}$ | 1162 | 0 | 1 | .42 | .49 |
| Hotel's Customer Service Evaluation Prior to SG $CSE\_HPRE_j$ | 81 | 24.81 | 34.20 | 30.199 | 2.21 |
| Hotel's Fraction of Guests on Vacation Trips Prior to SG $VAC\_HPRE_j$ | 81 | .155 | .784 | .410 | .140 |
| Hotel's Fraction of Guests on Business Trips prior to SG $BUS\_HPRE_j$ | 81 | .075 | .717 | .460 | .148 |
| Hotel's Customers' Average Share of Stays prior to SG $BRANDLOYAL\_HPRE_j$ | 81 | .282 | .458 | .352 | .003 |

Table 3: OLS Regression at Hotel-Time

| Dependent Variable: Customer Service Evaluation $CSE_{jt}$ | | |
|---|---|---|
| Independent Variable | Coefficient | Standard Error |
| Intercept | −6.5 | .46 |
| Fraction of Customers on Business Trips $BUS_{jt}$ | 6.65 | .43 |
| Fraction of Customers on Vacation Trip Dummy $VAC_{jt}$ | 6.70 | .45 |
| Average Customer's Share of Stays $BRANDLOYAL_{jt}$ | 1.97 | .74 |
| Months post SG $TIME_{jt}$ | −.00015 | .00 |
| SG Program Status Dummy $SG_{jt}$ | .0046 | .13 |

Table 4: Model Estimates

| Variable | Coefficient | Estimate | Std. Err. |
|---|---|---|---|
| ***Fixed Effect*** | | | |
| Customer's Business Trip Dummy $BUS_{jti}$ | $\pi_1$ | .290** | .162 |
| Customer's Vacation Trip Dummy $VAC_{jti}$ | $\pi_2$ | .476* | .141 |
| Customer's Share of Stays $BRANDLOYAL_{jti}$ | $\pi_3$ | 4.035* | .183 |
| *MONTH1 (01/98)* | $\pi_4$ | .950* | .316 |
| *MONTH2 (02/98)* | $\pi_5$ | .621* | .272 |
| *MONTH3 (03/98)* | $\pi_6$ | .301 | .275 |
| *MONTH4 (04/98)* | $\pi_7$ | .332 | .275 |
| *MONTH5 (05/98)* | $\pi_8$ | .109 | .267 |
| *MONTH6 (06/98)* | $\pi_9$ | .452 | .278 |
| *MONTH7 (07/98)* | $\pi_{10}$ | .122 | .243 |
| *MONTH8 (08/98)* | $\pi_{11}$ | −.053 | .294 |
| *MONTH9 (09/98)* | $\pi_{12}$ | .299 | .239 |
| *MONTH10 (10/98)* | $\pi_{13}$ | −.202 | .275 |
| *MONTH11 (11/98)* | $\pi_{14}$ | .480* | .227 |
| *MONTH12 (12/98)* | $\pi_{15}$ | 1.183* | .272 |
| *MONTH13 (01/99)* | $\pi_{16}$ | −6.229* | .430 |
| *MONTH14 (02/99)* | $\pi_{17}$ | −2.120* | .325 |
| *MONTH15 (03/99)* | $\pi_{18}$ | .227 | .213 |
| Hotel Grand Mean Intercept Term—Eq 3 | $\gamma_{000}$ | 28.409* | .355 |
| SG Program Effect Mean Intercept Term—Eq 4 | $\gamma_{010}$ | .138 | .168 |
| Hotel's Customer Service Evaluation Prior to SG $CSE\_HPRE_j$ | $\gamma_{011}$ | .162* | .050 |
| Hotel's Fraction of Guests on Business Trips prior to SG $BUS\_HPRE_j$ | $\gamma_{012}$ | 7.706* | 3.100 |
| Hotel's Fraction of Guests on Vacation Trips prior to SG $VAC\_HPRE_j$ | $\gamma_{013}$ | 8.194* | 3.344 |
| Hotel's Customers' Average Share of Stays prior to SG $BRANDLOYAL\_HPRE_j$ | $\gamma_{014}$ | 1.302 | 2.736 |
| Standard Deviations of Random Effects | | | |
| SD($e_{jti}$) | $\sigma$ | 7.853* | .1127 |
| SD($r_{0jt}$) | $\rho$ | .962 | .6669 |

| | | | |
|---|---|---|---|
| SD($u_{00j}$) | $\tau_{00}$ | 1.977* | .2707 |
| SD($u_{01j}$) | $\tau_{01}$ | .436* | .1721 |

Notes:
*CSE_HPRE*$_j$, *BUS_HPRE*$_j$, *VAC_HPRE*$_j$, and *BRANDLOYAL_HPRE*$_j$ are mean centered.
(*): Significant at 5% level
(**): Significant at 10% level

Table 5: Posterior Means of SG Program Effects ($\overline{\beta}_{01j}$)

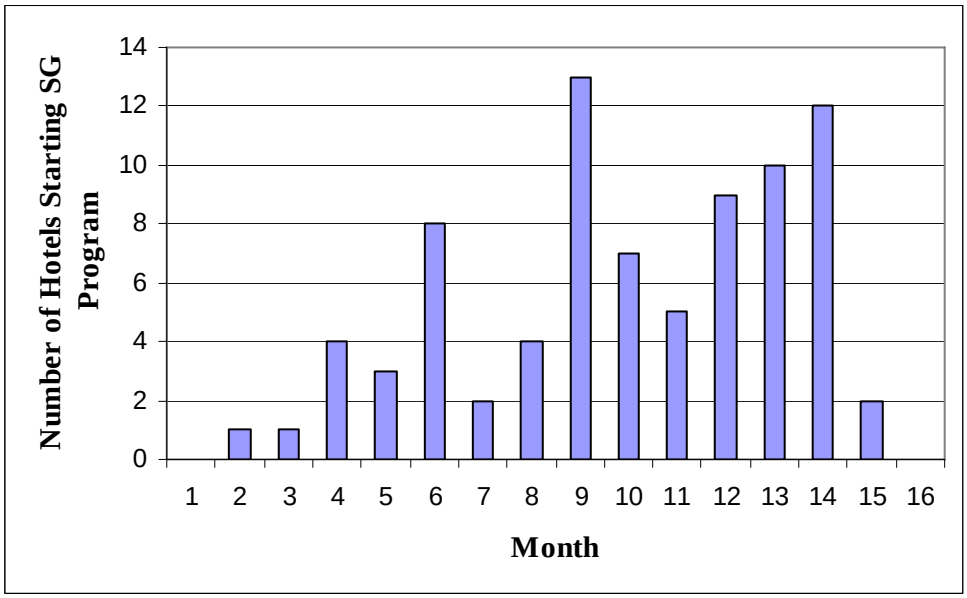
| | |
|---|---|
| Number of observations | 81 |
| Mean | .138 |
| Standard deviation | .356 |
| Minimum | −.990 |
| Maximum | 1.166 |

Table 6: Program Decisions versus Reward Decisions

| Dependent Variable: ($P_1 - P_0$) | | |
|---|---|---|
| Independent Variable | Coefficient | Standard Error |
| $\alpha$ | .94* | .016 |
| $\delta$ | −.066* | .031 |

(*): Significant at 5% level

Figure 1: Distribution of Program Start Dates




Note: Month 1 is January 1998.

Figure 2: Distribution of Survey Points

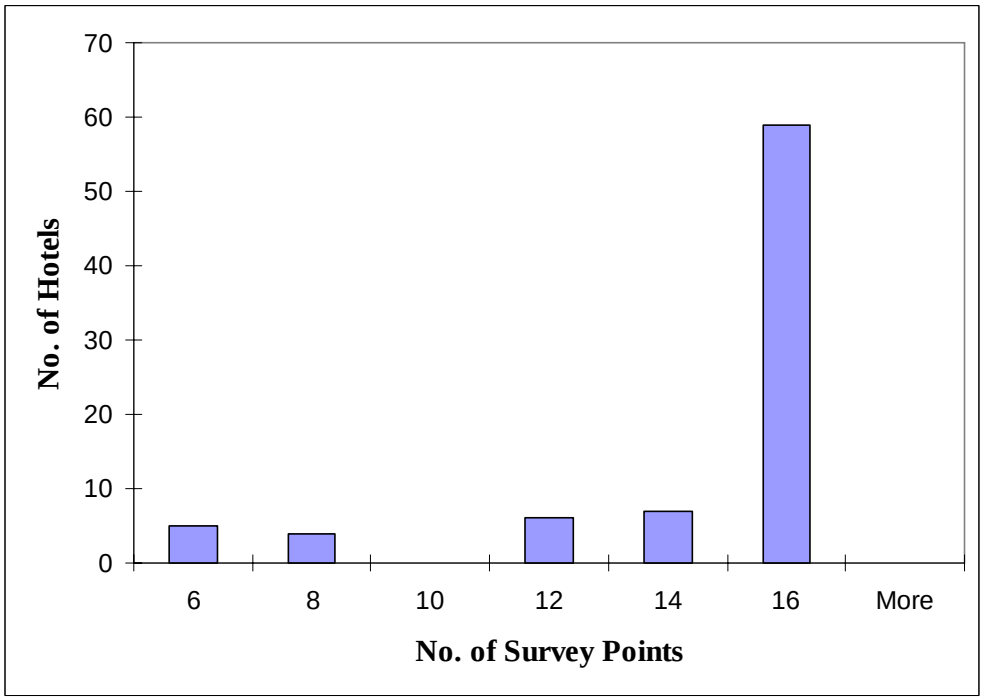


Figure 3: Surveys before and after Program Start Date

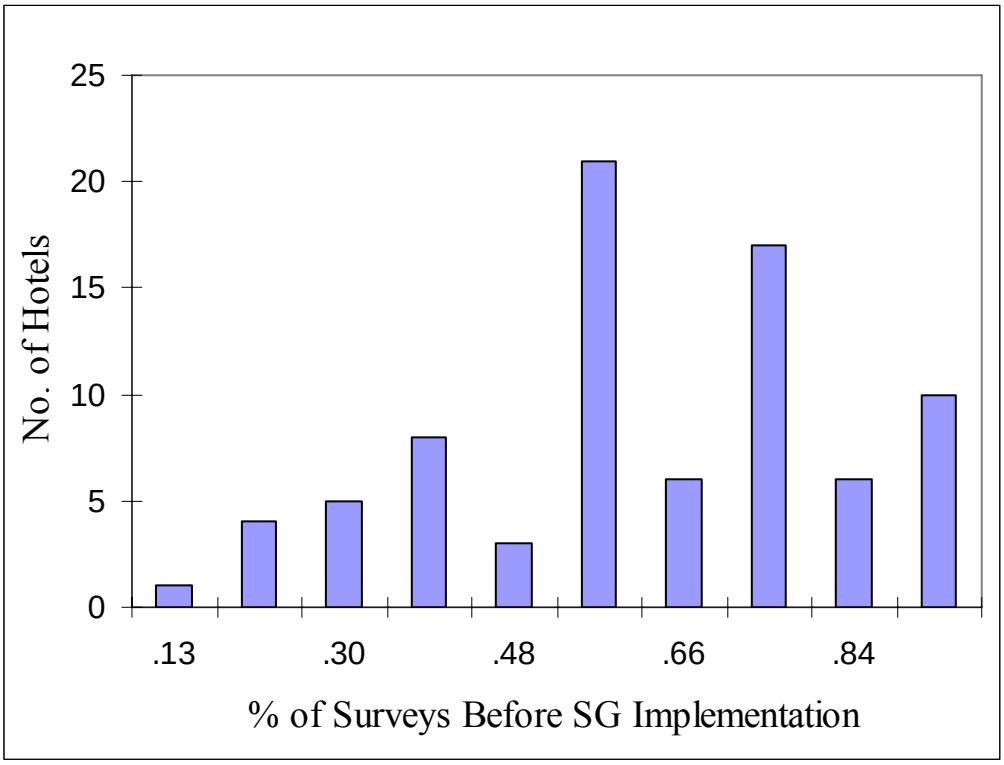



Figure 4: Scree Plot of CSE Factors

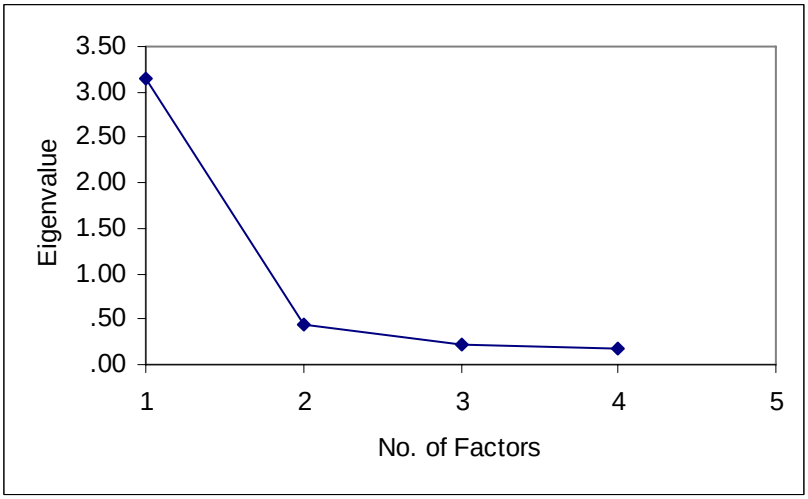
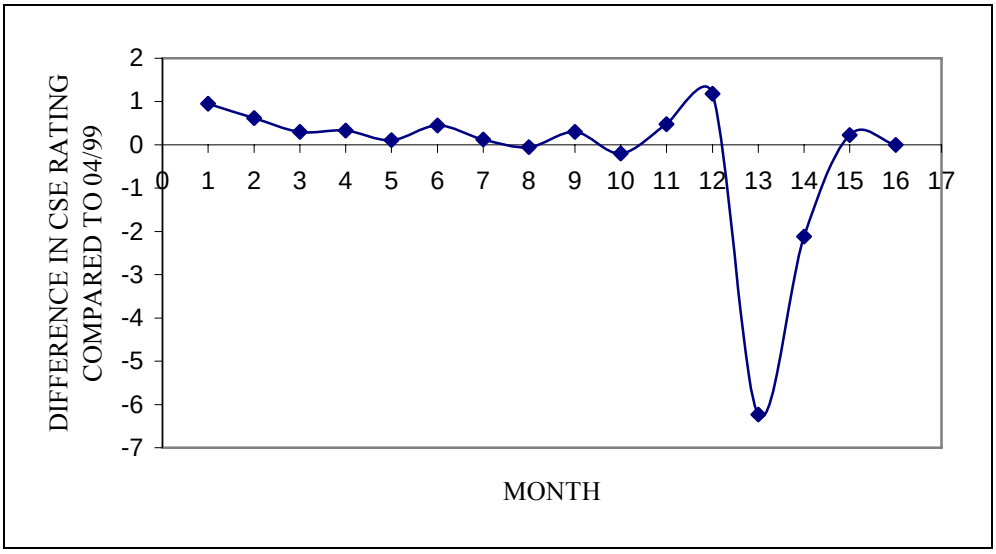
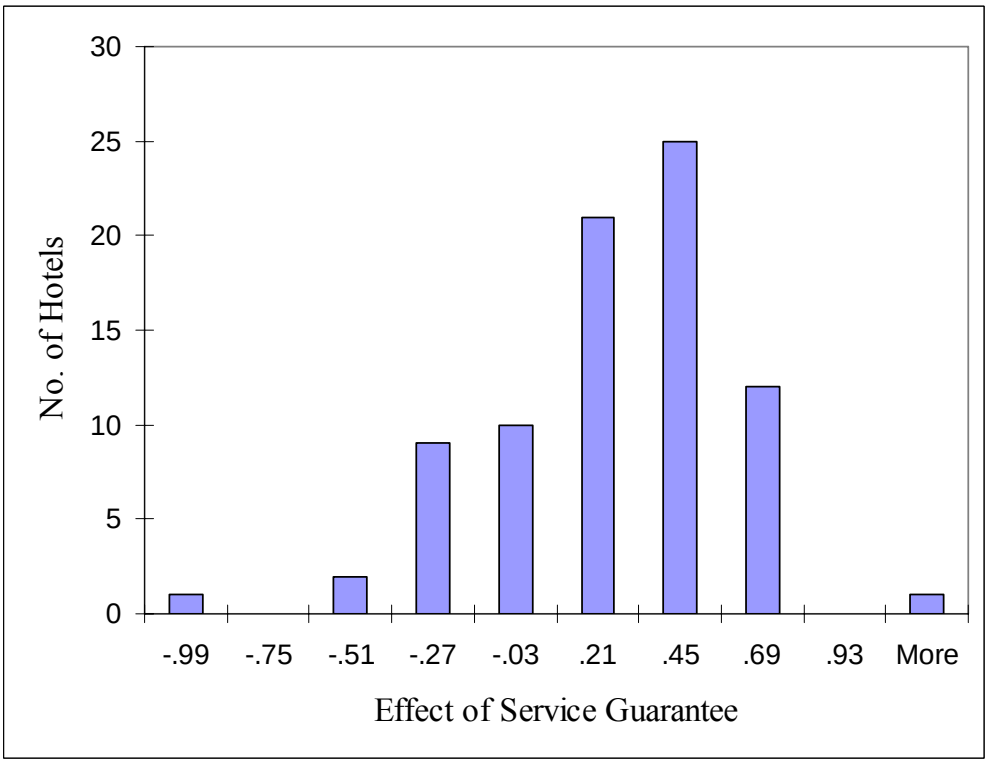


Figure 5: Month Effects on CSE



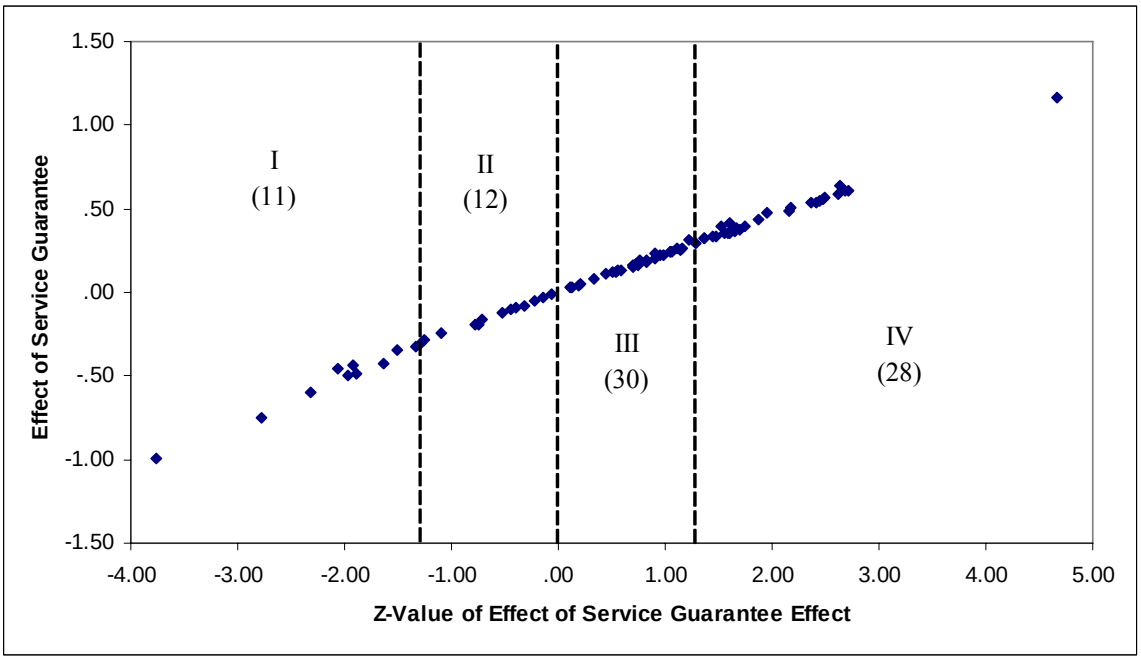

Note: Month 1 is 01/1998.

Figure 6: SG Program Effect



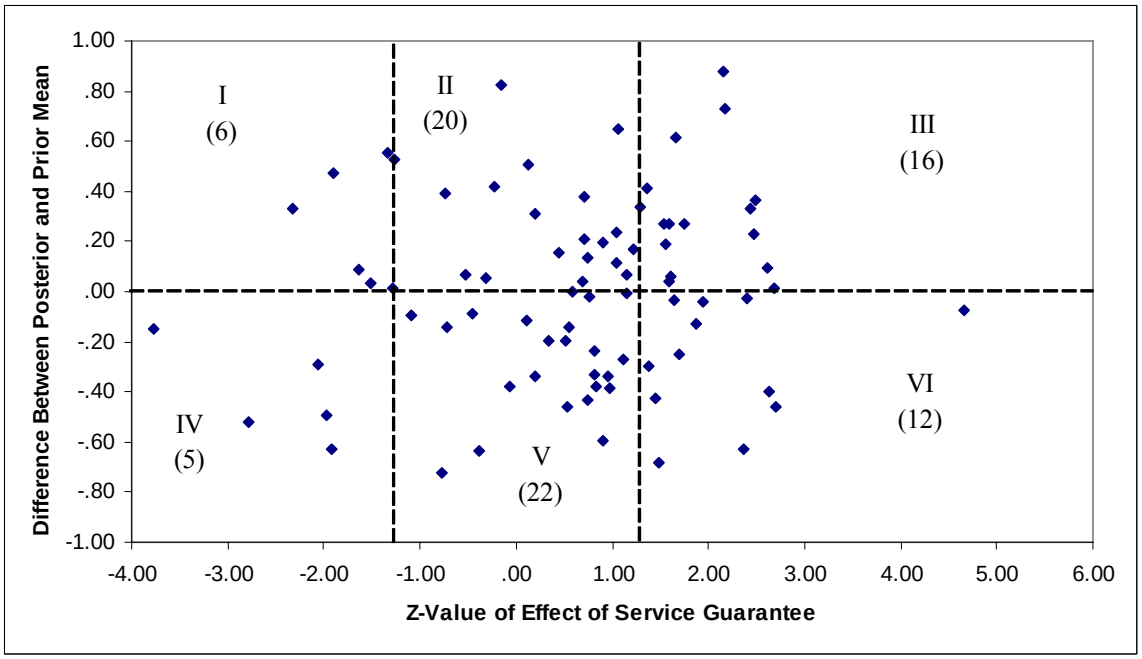

Note: X axis plots $\bar{\beta}_{01j}$

Figure 7: Statistical Significance of Program Effect




Notes: 1. Hotels in Areas 1 and 4 show significant SG effects.
2. The lines at –1.28 and 1.28 represent the 90% level of confidence.

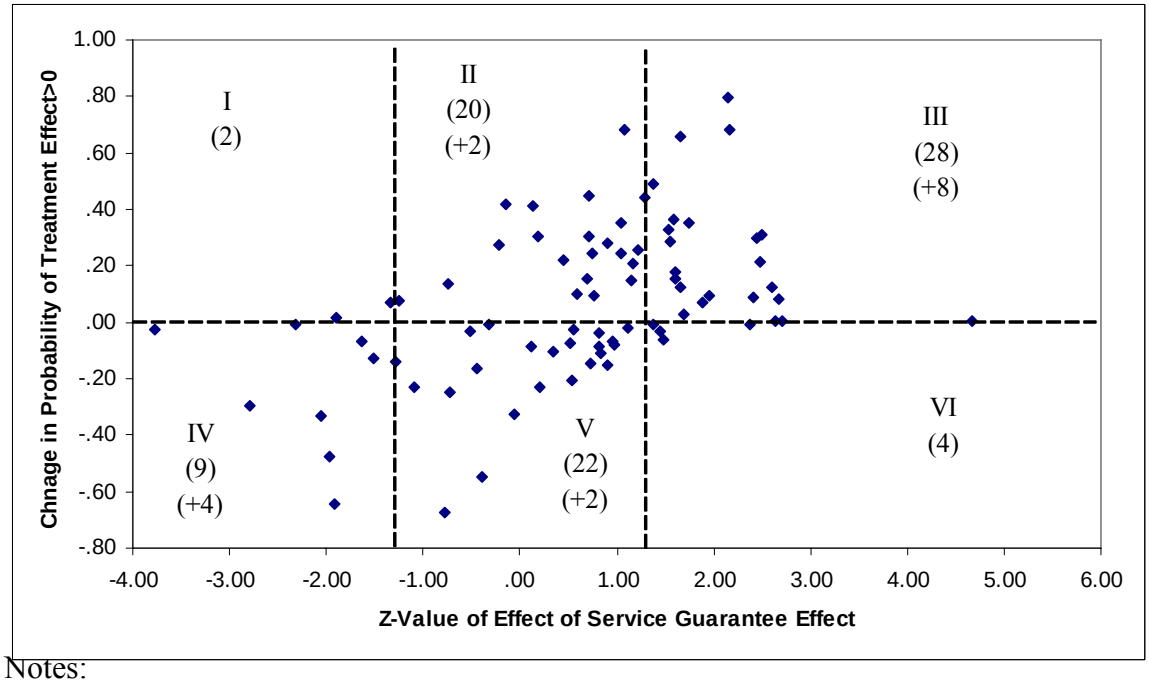
Figure 8: Reward Policy I




Notes: 1. X axis plots $\left(\overline{\beta}_{01j} / \lambda\right)$; Y axis plots $\overline{\beta}_{01j} - \tilde{\beta}_{01j}$.

2. Lines on x axis at –1.28, 1.28 represent the 90% level of confidence..
3. Numbers in parentheses show number of hotels.
4. Areas 1, II, and III are rewarded for exceeding expectations.

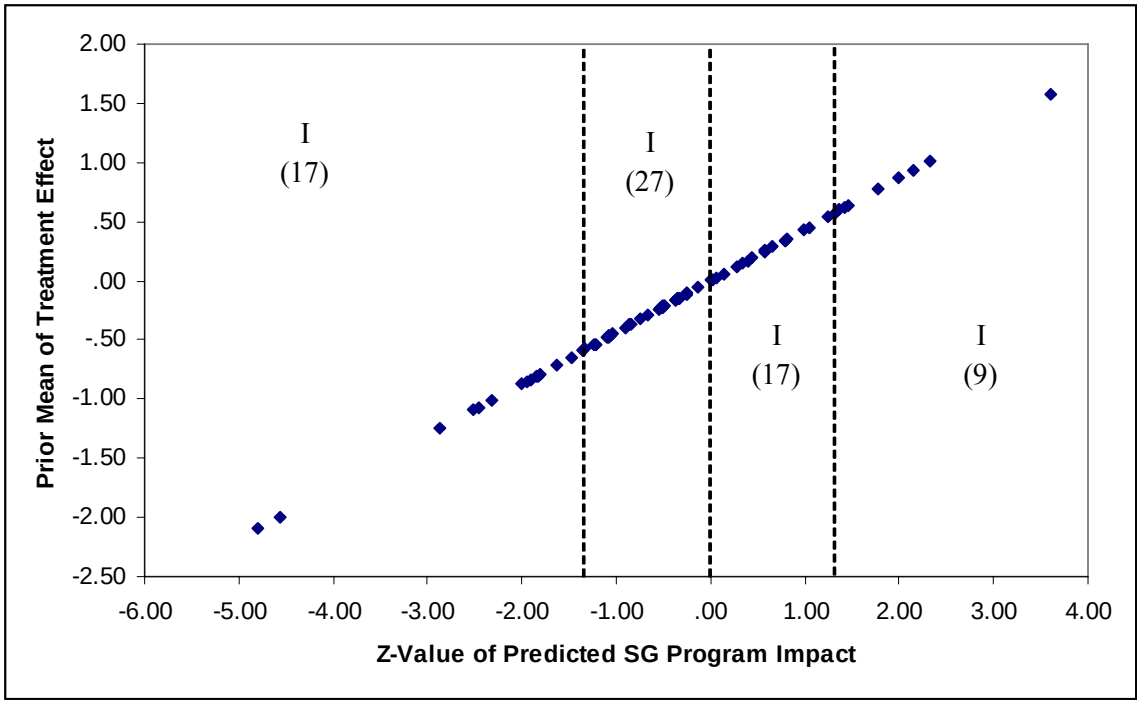
Figure 9: Reward Policy II




Notes:

1. X axis plots $\left(\overline{\beta}_{01j} / \lambda\right)$; Y axis plots $(P_0 - P_1)$.
2. Vertical lines on x axis at –1.28 and 1.28 represent the 90% level of confidence.
3. Areas 1, II, III are rewarded for exceeding expectations.
4. Numbers in first parenthesis indicate number of hotels in this area. The second parenthesis indicates the number of hotels that move into this area compared with Figure 8.

Figure 10: Targeting Hotels for Program Rollout




1. X axis plots $\left(\widetilde{\beta}_{01j} / \tau_{01}\right)$; Y axis plots $\left(\widetilde{\beta}_{01j}\right)$. Hotels in Area 1 are the priority targets.
2. The reference lines at –1.28 and 1.28 represent the 90% level of confidence.