The Scope and Persistence of Mere-Measurement Effects: Evidence from a Field Study of Customer Satisfaction Measurement

UTPAL M. DHOLAKIA
VICKI G. MORWITZ*

Self-generated validity research has demonstrated that responding to survey questions changes subsequently measured judgments and behavior. We examine the scope and persistence of the effect of measuring satisfaction on customer behavior over time. In a field experiment conducted in a financial services setting, we hypothesize and find that measuring satisfaction (*a*) changes one-time purchase behavior, (*b*) changes relational customer behaviors (likelihood of defection, aggregate product use, and profitability), and (*c*) results in effects that increase for months afterward and persist even a year later. These results raise questions concerning the design, interpretation, and ethics in the conduct of applied marketing research studies.

Firms routinely conduct customer surveys to determine how satisfied they are with the firm and its offerings and how likely they are to buy new product offerings, and to evaluate different aspects of the firm's marketing mix. An important assumption underlying all such surveys is that existing opinions are elicited from customers and that these do not influence their subsequent opinions or behaviors. Indeed, the American Marketing Association (AMA) code of ethics clearly segregates the conduct of marketing research from any form of sales or opinion-influencing activity. However, research on self-generated validity theory suggests that when responding to surveys, respondents are often induced by the measurement process to form judgments that would otherwise not be formed, which in turn influences subsequent responses and behaviors, making them more consistent with the expressed judgments (Feldman and Lynch 1988; Simmons, Bickart, and Lynch 1993). Moreover, such measurement-induced judgments are especially likely in contexts such as satisfaction or purchase intention surveys, in which most respondents are unlikely to have formed these judgments spontaneously beforehand or, indeed, given the issue much prior thought (Kardes 1988; Weiner 1985).

A related stream of research has shown that measurement-induced judgments, specifically the elicitation of behavioral intentions, can change respondents' subsequent actions. For example, Sherman (1980) showed that asking people to predict whether they would engage in a socially desirable (undesirable) behavior leads to increased (decreased) participation rates relative to people who are not asked to make such a prediction. In the marketing literature, several studies have shown that the process of measuring purchase intentions changes subsequent purchase behaviors (see Morwitz and Fitzsimons [2000] for a recent review), a phenomenon that has been called the mere-measurement effect (Morwitz, Johnson, and Schmittlein 1993) and the self-prophecy effect (Spangenberg and Greenwald 1999).

Two aspects of current knowledge regarding self-generated validity are important to note. First, most studies have defined these effects fairly narrowly. For example, studies on the mere-measurement effect have focused primarily on the influence of measuring intentions on one-time purchases (e.g., Fitzsimons and Williams 2000; Morwitz et al. 1993). Second, such effects are conceptualized in fairly short temporal terms, considering outcomes occurring instantaneously or shortly (generally within minutes) after the time when the judgment is measured (e.g., Fitzsimons and Shiv 2001; Morwitz and Fitzsimons 2000; Spangenberg and Greenwald 1999). While a few studies (e.g., Fitzsimons and Morwitz 1996; Morwitz et al. 1993) have shown effects over a longer time horizon (i.e., changes in purchasing within a six-month window after intentions measurement), none has explicitly

*Utpal M. Dholakia is assistant professor of management at the Jesse H. Jones Graduate School of Management, Rice University, Houston, TX 77005-1892 (dholakia@rice.edu). Vicki G. Morwitz is associate professor of marketing and the Edythe and George Heyman Research Fellow at the Stern School of Business, New York University, New York, NY 10012 (vmorwitz@stern.nyu.edu). The authors thank the financial services firm for sponsoring the study and Rick Bagozzi, Bob Westbrook, the editor, associate editor, and three reviewers for helpful comments made on earlier versions of this article.

examined the duration of self-generated validity effects. For example, it is not clear in the Morwitz et al. (1993) study whether the increased purchasing associated with intent measurement occurred the day after intentions were measured, the week after, or throughout the six-month interval.

This research describes the results of a large-scale field experiment designed to examine the scope and persistence of measurement-induced satisfaction judgments on subsequent behavior. We build on past research in two important ways. First, we expand the focus from just one-time purchase to multiple behaviors. In keeping with the increasing research interest in relationship marketing (e.g., Reinartz and Kumar 2000), we hypothesize and demonstrate that in addition to changing transactional behavior, measuring satisfaction changes subsequent relational behaviors (e.g., defection, total account ownership, and profitability) of customers.

Second, and more important, we examine the duration of measurement effects over a one-year period following satisfaction measurement. These effects are found to be durable, showing that although measuring satisfaction has an immediate effect on behavior, its maximum impact occurs several months after the survey, and this influence continues to persist even at the end of the one-year period.

The results of the field experiment show that the effect of measurement-induced judgments is persistent over time and broad in scope, suggesting that it should be considered carefully during the employment, conduct, and interpretation of survey-based marketing research studies. Specifically, these results suggest that researchers need to make adjustments when predicting the behavior of larger groups using such measures, and point to the need for procedures to minimize the distorting effects of measurement. The extent of influence uncovered here also raises questions of ethics when conducting applied marketing research, especially since selling under the guise of conducting research (sugging) violates the AMA code of ethics; indeed, legislation in the mid-1990s has outlawed this practice (Bowers 1995).

CONCEPTUAL OVERVIEW AND RESEARCH HYPOTHESES

The primary idea underlying self-generated validity theory is that when participants respond to surveys, only some of the elicited responses exist in the participant's memory. In all other cases, such judgments are generated by the participant from other available inputs and are therefore measurement induced. Consider the measurement of customer satisfaction, a very common type of judgment collected in applied marketing research. For most people, responses to questions in satisfaction surveys are likely to be measurement induced rather than available beforehand or spontaneously formed. This is because explicit directed thoughts regarding the satisfaction level with a particular product are unlikely to occur spontaneously for most respondents in the absence of specific questioning or a striking (positive or negative) experience with the product (Hastie 1984; Schul and Schiff 1993; Schuman and Presser 1980).

Measurement-induced judgments influence both subsequent judgments, as evidenced by carryover effects in surveys in which respondents give answers that are consistent with inputs rendered accessible by previous responses (Bickart 1993; Fitzsimons and Shiv 2001; Tourangeau and Rasinski 1988), and behavior, as evidenced by the mere-measurement effect (e.g., Morwitz et al. 1993) and self-prophecy effect research (Spangenberg and Greenwald 1999). An important reason for this greater influence of measurement-induced judgments relative to retrieved judgments is that they bring the respondent's normally automatic behavior under conscious volitional control, increasing awareness of other, more appropriate responses or courses of action in light of the expressed opinion (Feldman and Lynch 1988). Behaviors that were habitual before now become deliberate.

Even when behavior is temporally separated from measurement, the influence of measurement-induced judgments may be greater if and when the causal analyses underlying measurement-induced judgments are more accessible relative to the bases of automatic judgments, making the former more reliable as a basis of action. Psychological processes, such as script evocation (Sherman 1980), and increased accessibility of the judgment (Kardes et al. 1993) may play an important role in bringing about this influence (see Spangenberg and Greenwald [1999] for a detailed discussion). Responding to a satisfaction survey should therefore result in not only the formation of evaluations regarding the firm and its offerings for many participants but also subsequent behaviors consistent with these expressed evaluations. Feldman and Lynch (1988, p. 423) summarize this expected effect in the context of a job-satisfaction survey as follows: "The entire sequence of questioning serves as the basis for attitude formation, which then serves as the most salient basis for the development of an intention" and presumably results in behavior.

Moreover, the direction of the behavioral change is likely to depend on the valence of expressed satisfaction in the surveyed group. In competitive markets, existing customers tend to have high levels of satisfaction with their current providers (otherwise they would defect; see Fornell [1992] for a detailed discussion). We therefore expect the direction of behavioral change to be generally positive at the firm level such that the satisfied customers should engage in greater purchase of the firm's products following satisfaction measurement. However, the opposite should occur in the case of dissatisfied customers. They should purchase fewer of the firm's products after the survey. On the basis of this discussion, we expect the following behavior-influencing role of satisfaction judgments:

> **H1:** The likelihood of future purchase behaviors will be higher (lower) when satisfied (dissatisfied) customers respond to a satisfaction survey than when they do not.

We note that this hypothesis extends the scope of mere-measurement effects from behavioral intentions in the extant literature to customer satisfaction judgments. Also, in the

financial services case, product purchase corresponds to the opening of a new account by the customer.

In the increasingly influential paradigm of relational marketing, there is recognition that customers' relational behaviors—that is, on-going purchase and use of a portfolio of offerings—are of great consequence (Reichheld 1996; Reinartz and Kumar 2000). Perhaps the most important relational measure is the length of relationship between the customer and the firm (Reichheld 1996; Sheth and Parvatiyar 1995), which is operationalized by defection rate, the proportion of customers that defect from the firm during any given time period (Reinartz and Kumar 2000). Comprehensive customer databases available in many industries make defection rates widely available for relationship-marketing activities. A second important relational measure is the aggregate purchase of the firm's offerings during any time period. Restaurants or catalog retailers measure this by the total amount spent by the customer in a month while in the financial-services setting; this is measured by the customer's total number of accounts. Finally, a third important relational measure is the customer's profitability to the firm during a given time period.

Relational measures such as defection rates, aggregate product purchase, and customer profitability are more indicative of firm performance than any transactional measure (Sheth and Parvatiyar 1995). To date, perhaps because of the narrow definition, research on self-generated validity and, more specifically, mere-measurement effects has given little attention to relational behaviors. In keeping with our interest in understanding the scope of such effects, we examine how satisfaction judgments influence such relational behaviors in addition to transactional behaviors.

Most applied satisfaction studies elicit both specific judgments (e.g., ambiance of store, friendliness of staff) as well as overall satisfaction with the firm. According to self-generated validity theory, by making these different specific and general judgments more accessible to the respondent, satisfaction measurement should ultimately influence a variety of behaviors pertaining to the firm. This is referred to in the attitude literature as the "principle of compatibility," in which general accessible attitudes show substantial correlations with behavior if the behavioral measures are aggregated across a number of specific behaviors (Eagly and Chaiken 1993). Further, responses to earlier specific questions increase the accessibility of positive information for the satisfied customers, resulting in higher satisfaction ratings for later questions and a more positive overall evaluation in the end (Schul and Schiff 1993). This positive judgment in turn has been shown to influence satisfaction consequences such as loyalty and profitability (Fornell 1992) and may do so through reduced defection and greater product use at the individual customer level. The following hypothesis formalizes this discussion:

H2: The likelihood of future relational behaviors will be higher (lower) when satisfied (dissatisfied) customers respond to a satisfaction survey than when they do not.

Hypothesis 2 extends the scope of mere-measurement effects from one-time purchase to relational customer behaviors.

Little research to date has explicitly examined the timing or duration of self-generated validity effects. In considering the role of time, Feldman and Lynch (1988) argue that the probability that a response to one survey item will be retrieved and will influence a subsequent item is inversely related to the time between the two items. This suggests that the largest effects of satisfaction measurement should occur soon after measurement and then decay over time.

We have argued that the process of eliciting satisfaction forms these judgments for many respondents for whom none existed before. Moreover, many customers may perceive the process of participation in a satisfaction survey to be of value. In such cases, there is likely to be some positivity bias in which the judgments may become even more positive on account of participation—that is, "since they care enough to ask, they must be really good." Finally, for many participants, responding to a battery of specific satisfaction measures may actually have knowledge value, increasing awareness about the variety of the firm's offerings and encouraging future purchases. For all of these reasons, we expect the decay rate of the cognitive effect to be slow. Measuring satisfaction should result in a persisting increase in positive feelings toward the firm for satisfied customers.

Considerable research also suggests that such judgments, even when very positive, alone do not provide sufficient impetus to engage in behavior, such as to purchase a new product or to significantly increase product use (Bagozzi and Dholakia 1999). For financial products, such behavior also requires some energizing goal or event to occur, such as a change in job, marriage, new birth, inheritance, and so on. In the same way, defection from a financial services firm may occur because the respondent learns about a new provider with better options, because of a move to another city, and so on. Such energizing events correspond to what social psychologists have called "channel factors," response pathways that serve to elicit or sustain behavioral intentions and behavior with a particular intensity or stability over an extended period of time (Ross and Nisbett 1991). Assuming that such events occur randomly over time following the survey, we expect the cumulative probability of customer behaviors, such as purchasing a new product, to increase over time as the cumulative probability of behavior-inducing events increases. Along similar lines, we expect customers' opportunity and ability with regard to discretionary time and money to fluctuate as well, with the cumulative probability of favorable opportunity and ability states increasing over time.

Moreover, research on associative memory (e.g., Gillund and Shiffrin 1984) suggests that if the measurement-induced judgment is retrieved at such future action points, its accessibility may actually be augmented each time, which serves to extend its influence. On the whole then, we expect the cognitive effect of satisfaction measurement to persist long after the survey.

Thus, while the cognitive influence of measuring satis-

faction should be the greatest right after measurement and then decay slowly over time, the cumulative probability of behavior-inducing events should increase over time. Together, these processes should result in an increasing influence of measuring satisfaction on firm-related customer behaviors over time, with the greatest influence occurring at some interval after satisfaction measurement and then subsequently declining over time. This discussion is summarized in the following hypothesis:

H3: The effect of measuring satisfaction on customer behaviors will persist, increasing from the time of measurement, reaching a maximum at some interval following measurement, and then decreasing over time.

STUDY DESIGN

A large-scale field experiment was conducted to test the hypotheses. A two-phase pretest-posttest experimental design (Maris 1998) was used. This design has two important advantages over the posttest-only design used in the extant mere-measurement effect research: (*a*) it permits error variance caused by consistent individual differences to be removed, thereby increasing power; and (*b*) it permits the groups to be equated for baseline differences, thereby increasing the internal validity of the design. Retail customer households enrolled in the client management program of a large financial services firm were randomly assigned to either the experimental or the control group. Households in the experimental group participated in a customer satisfaction study that asked them to evaluate and rate the value of specific features of the client management program. The survey was conducted by telephone and solicited participation from the "person in your household who makes most of the decisions about which financial services you use and where you conduct your financial business."

The survey took between 10 and 12 minutes to complete on average and asked a number of satisfaction questions eliciting respondents' evaluation of specific program features (e.g., estate planning services, consolidation and monitoring of accounts, retirement planning services) followed by the general question, "Overall how satisfied are you with the (financial institution name)?" A seven-point "extremely dissatisfied–extremely satisfied" scale was used to obtain participant responses to all specific questions and the general question. A total of 945 customer households participated in this research and answered the telephone survey completely. To focus exclusively on the influence of satisfaction measurement on subsequent customer behaviors, the study asked no questions pertaining to future behavioral intentions.

A total of 1,064 customer households, randomly assigned to the control group, did not participate in the satisfaction study. The two groups were well matched on both age of primary head of household (60.4 years vs. 59.5 years, $p > .16$) and average monthly profitability ($107.70 vs. $101.70, $p > .4$). Households assigned to the experimental group had an average of 4.13 accounts with the firm prior to the study, compared with an average of 3.78 accounts for households assigned to the control group. This difference was small but statistically significant ($p < .01$). Including the baseline measure of total number of accounts provided virtually identical results in tests of hypotheses 1 and 2 below, suggesting that this initial difference between the two groups did not significantly influence the results. In general, the two groups were well matched on key measures prior to the experimental manipulation (i.e., customer satisfaction measurement).

Following completion of the satisfaction study, all of the households in both groups were withheld from any direct marketing activity for a period of one year. In other words, none of these 2,009 households participated in any subsequent marketing research study conducted by the financial institution or its affiliates, nor did they receive any special direct mail (except for monthly account statements and enclosed inserts that all customers of the firm received), telemarketing calls, or courtesy calls from firm personnel during this time period. This was done to prevent specific marketing activities or programs from confounding the results of the study.

Key customer performance metrics were collected from the firm's customer database at the start of the study (referred to as baseline measures) as well as one year later (referred to as one-year measures) for all households in both groups. The following measures were collected: (*a*) new purchase: whether the customer household opened any new accounts during the one-year time period and the month in which the first such new purchase was made; (*b*) total number of accounts: the total number of accounts with the firm at the start of the study (baseline total accounts) and one year later (one-year total accounts); the total number of accounts was computed for each customer household by adding all of the credit, deposit, and investment accounts with the firm in that particular month; (*c*) customer profitability: current monthly contribution to the firm at the start of the study (baseline profitability) and one year later (one-year profitability). This variable is computed by many financial institutions using a standard algorithm based on well-accepted activity-based cost accounting practices. With this algorithm, the customer's monthly contribution is computed as the difference between total revenues (including fees, interest income, service charges, etc.) and total costs (including interest expenses, servicing costs, transaction costs, etc.). The same algorithm was used to compute customer profitability for all customers included in the analysis (both groups) at both time periods; and (*d*) customer defection: whether the customer defected from the firm and the month of defection. A customer was defined as having defected from the firm if it closed all of its accounts at any time during the one-year time period of the study.

RESULTS

Hypothesis 1

We hypothesized that merely measuring satisfaction would increase purchase behavior for satisfied experimental

group customers and decrease purchase behavior for dissatisfied experimental group customers relative to a control group for whom satisfaction was not measured. The average overall satisfaction rating for participants in the experimental group was 5.93 (SD = 1.33), measured on a seven-point scale. Further, of the 945 participants, only 39 of the surveyed customers indicated a score of three or less, suggesting overall positive customer satisfaction with the firm. We therefore expect to see positive effects of satisfaction measurement on behavior in our tests of hypotheses below.

To test hypothesis 1, we conducted a logistic regression analysis with customer group as the categorical independent variable and the binary purchase variable as the dependent variable. The group variable was a strong predictor of purchase ($\beta = 1.91$, Waldstatistic = 295.4, $p < .001$), suggesting that experimental group customers were significantly more likely to make one or more purchases from the firm than were control group customers. Whereas 51% of the customers for whom satisfaction was measured made a new purchase, only 13.3% of the control group made a new purchase during the one-year time period, supporting hypothesis 1.

Hypothesis 2

We hypothesized that merely measuring satisfaction would result in significant differences in relational behaviors between the experimental and control groups. We considered the following relational behaviors in testing this hypothesis: (*a*) defection rate, (*b*) total number of accounts, and (*c*) customer profitability. The results for each of these relational measures are discussed separately.

Defection Rate. A logistic regression analysis was conducted with customer group as the categorical independent variable and the binary defection variable as the dependent variable. Results showed that the group binary independent variable was a strong negative predictor of defection ($\beta = -1.03$, Waldstatistic = 44.1, $p < .001$), indicating that customers in the experimental group were significantly less likely to defect. While 16.4% of the customers in the control group defected during the one-year time period, only 6.6% of the experimental group customers defected during this time, supporting hypothesis 2.

Total Number of Accounts. In this case, results of a 2 group (experimental, control) × 2 time (baseline, one-year) repeated measures ANOVA indicate that the effect of group was significant ($F(1, 1{,}757) = 60.02$, $p < .01$), as was the effect of time ($F(1, 1{,}757) = 21.13$, $p < .01$). The two-way group × time interaction was also significant ($F(1, 1{,}757) = 286.3$, $p < .01$). Paired-comparison *t*-tests showed that in contrast to the significant increase in total accounts for the experimental group ($M_{\text{baseline}} = 4.16$ vs. $M_{\text{one-year}} = 5.45$, $t(872) = 12.99$, $p < .001$), total accounts actually declined significantly for the control group after one year ($M_{\text{baseline}} = 4.12$ vs. $M_{\text{one-year}} = 3.39$, $t(887) = -10.89$, $p < .001$), after adjusting for defected customers from both groups. The results of this analysis provide support for hypothesis 2.

Customer Profitability. Here, too, the results of a 2 group × 2 time repeated measures ANOVA indicate that while the effect of group was not significant ($F(1, 1{,}770) = 1.66$, $p > .1$), the effect of time was significant ($F(1, 1{,}770) = 12.93$, $p < .01$). More important, the two-way group × time interaction was significant ($F(1, 1{,}770) = 10.67$, $p < .01$). Paired-comparison *t*-tests showed that after adjusting for defected customers, whereas the profitability of customers in the control group declined sharply over the year ($M_{\text{baseline}} = \$111$ vs. $M_{\text{one-year}} = \$97.20$, $t(889) = -4.372$, $p < .001$), for customers in the experimental group, it remained more or less unchanged ($M_{\text{baseline}} = \$109$ vs. $M_{\text{one-year}} = \$107.80$, $t(883) = -.27$, NS). The results show that customers for whom satisfaction is measured show a better profitability profile than comparable customers for whom it is not measured. Note, however, that profitability did not increase for the experimental group in this case, in contrast to the other dependent measures. On the whole, hypothesis 2 is supported for all three relational measures considered.

Hypothesis 3

Given our focus on timing in hypothesis 3, we consider the two customer behaviors in which timing is relevant: new purchase and defection. For customers engaging in these behaviors, the month in which it first occurred during the one-year period was available. These data allow us to examine whether measuring satisfaction affects the risk or hazard of occurrence of the behavior and how the profile of the hazard changes over the duration of the experiment.

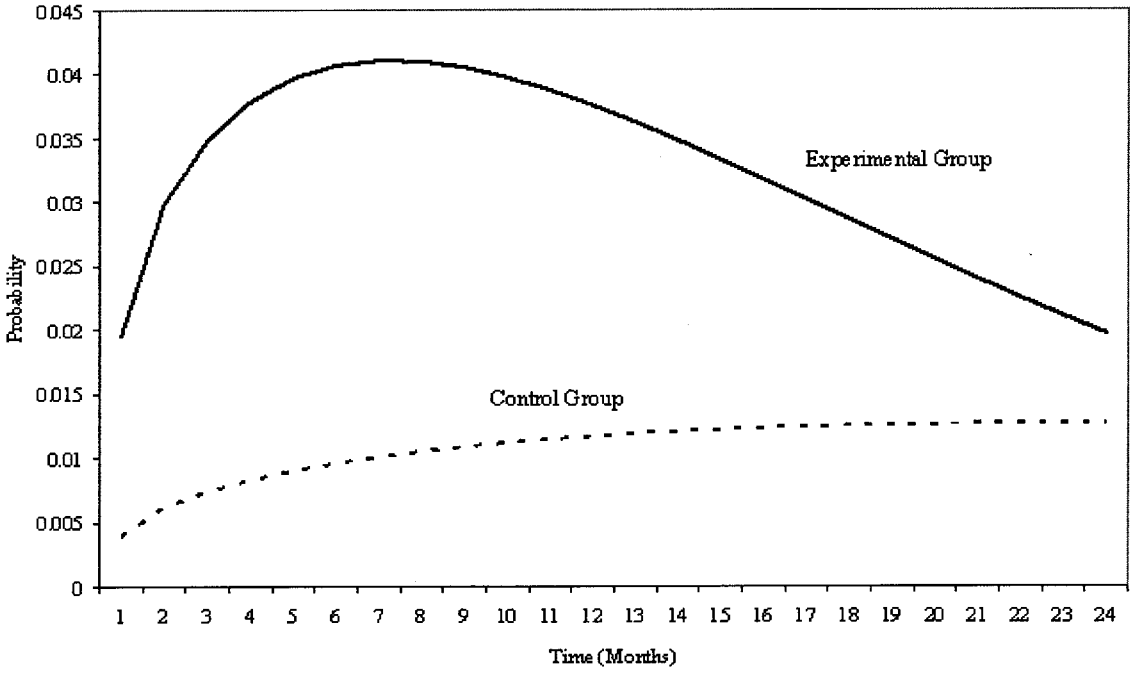
SAS PROC LIFEREG was used to fit parametric duration models, using data on the number of months from the time of satisfaction measurement until the new purchase or defection. It is important to note that both duration data series are right censored. In other words, we only know if a customer defected or made a new purchase over the duration of the experiment (i.e., within 12 months), but any transactions occurring after this point are not observed. The LIFEREG procedure is designed to handle such right-censored data. This procedure was used to estimate the following model:

$$T_i = \exp(\beta_0 + \beta_1 x_i + \sigma\varepsilon_i),$$

where for subject i, T_i is a random variable denoting the time of the event (either adding a new account or defecting), ε_i is a random error term, x_i is a dummy variable indicating whether or not subject i's satisfaction was measured, and β_0, β_1, and σ (the scale parameter) are parameters that are estimated using a maximum likelihood procedure. The LIFEREG procedure allows five different distributions for the error term ε_i: one-parameter extreme value, two-parameter extreme value, normal, logistic, and log-γ. These in turn lead to the following distributions for T_i, each of which

FIGURE 1

NONCUMULATIVE PROBABILITY OF NEW PURCHASE BY GROUP OVER TIME



makes different assumptions about the hazard function: exponential (assumes a constant hazard rate), Weibull (allows the hazard rate to increase or decrease over time), log-normal (assumes the hazard has an inverted U shape), log-logistic (assumes the hazard has an inverted U shape or declines over time), and γ (allows a wide variety of shapes for the hazard function, including those of the previous models and a U-shaped function). Since many of these models are nested within others, likelihood ratio tests may be used to determine the best-fitting model (see Allison [1995], pp. 88–90, for details on conducting goodness-of-fit tests using likelihood-ratio statistics). The LIFEREG procedure was used to estimate the five variants of duration models (i.e., the five different error term assumptions) for both new purchase and defection. Likelihood-ratio tests indicated that the Weibull model fitted best in both cases.

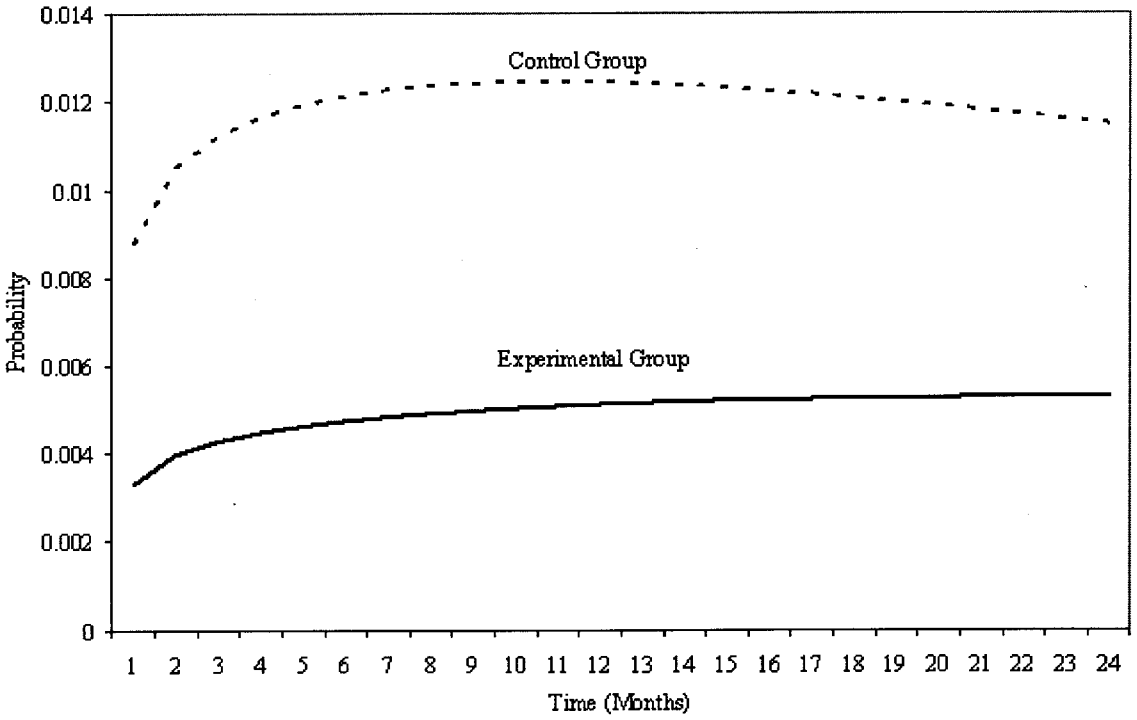
New Purchase. The model estimates for the Weibull new purchase model indicate that customers whose satisfaction was measured had significantly shorter times until purchase than those for whom satisfaction was not measured ($\beta_1 = -1.18$, $\chi^2(1) = 213.51$, $p < .001$). We note that similar results were obtained for all five duration models. An estimate for the ratio of the average time until new purchase across groups, $e^{\beta 1}$, indicated that the expected time to purchase for those whose satisfaction was not measured was 224% greater than for those whose satisfaction was measured. The scale parameter was between .5 and 1 ($\sigma = .73$), indicating that the hazard of defection increases over time at a decreasing rate. Figure 1 shows the noncumulative probability of purchasing for the two groups predicted by the Weibull new purchase model.

These results support hypothesis 3 for new purchase, showing that the positive effect of measuring satisfaction on purchasing increases over the first six months and then, although still large, decreases in magnitude thereafter. This model also predicted that the effect of measuring satisfaction on new purchase will persist for 31 months after the survey, though this prediction is based on extrapolating far beyond the observed data and should be interpreted with caution.

Defection. The results for the defection data were similar. Results of the Weibull defection model indicated that customers whose satisfaction was measured had significantly longer times until defection than those for whom satisfaction was not measured ($\beta_1 = .86$, $\chi^2(1) = 37.66$, $p < .0001$). As with the new purchase data, this pattern of results also held for the other four duration model types. Results suggest that the expected time to defection for those whose satisfaction was measured is 135% greater than for those whose satisfaction was not measured and that the hazard of defection increases at a decreasing rate ($\sigma = .87$). Figure 2 shows the noncumulative probability of defection for the two groups predicted by the Weibull defection model.

FIGURE 2

NONCUMULATIVE PROBABILITY OF DEFECTION BY GROUP OVER TIME



Consistent with hypothesis 3, these results show that the effect of satisfaction measurement on defection increases for eight months after the survey and then, although still large, reduces in magnitude. The model also predicted that measuring satisfaction would continue to have an influence on reducing defection for 93 months; again, however, this result should be interpreted with caution. In general, results for hypothesis 3 show that the effects resulting from satisfaction measurement are persistent, increasing for several months after the measurement, before starting to decline.

GENERAL DISCUSSION

The results of the field experiment show that measuring satisfaction influences not just one-time purchase but also relational behaviors of customers over an extended period of time. Even more important, the influence of satisfaction measurement on respondents' behaviors is found to increase for several months after judgment elicitation and persists even a year later.

Building on the existing knowledge regarding self-generated validity theory, we argued that the formation of measurement-induced judgments resulting from satisfaction measurement, the increased positivity of judgments on account of perceived participation and knowledge value, and the self-generated validity of the overall evaluation on account of earlier specific evaluations may all influence the customer's behaviors with regard to the firm in a stable and sustained fashion. Another possibility is that the measurement process leads respondents to temper their forced judgments away from self-perceived extremes. However, the true behavioral propensity remains unchanged and is revealed in behavior.

Finally, the observed durability of these effects may occur from channel factors, such as energizing life events, from the augmented accessibility following retrieval on subsequent occasions (Gillund and Shiffrin 1984), or from a combination of the two mechanisms. However, the field-experimental methodology, the naturalistic setting, and the applied context of the research did not allow us to collect process measures to fully validate these accounts or to determine the specific mechanisms underlying these observed effects. It is therefore important that future studies corroborate these results with greater internal validity through a laboratory experiment methodology and appropriate process measures. In the same way, this research was limited to highly satisfied customers of a single firm in one industry. It is possible that these results may be subject to the vicissitudes pertaining to the firm itself or to the financial services industry more generally. Similar studies need to be undertaken across firms

and in other industries to replicate our findings and to provide generalizability.

Finally, it is also important to point out that the participants of our study had a formal ongoing relationship with the firm and were among its more profitable and higher potential customers. These factors may have magnified the positive effects of satisfaction measurement on participants. Such effects may likely be mitigated in the following cases: (*a*) for buyer-seller relationships consisting largely of discrete transactions or having relatively low switching costs and (*b*) those instances where customers have less volitional control over their purchases because of limited opportunity or ability. These issues warrant future research attention.

Most applied research studies elicit both specific and general evaluations of satisfaction. Prior research on measurement-induced satisfaction judgments (e.g., Schul and Schiff 1993) has shown that the order in which these measures are presented influences the valence of responses and, presumably, subsequent behavior. When specific judgments are provided first followed by the overall judgment, the earlier specific judgments influence the later overall judgment, regardless of valence. Thus, customers satisfied with specific features may show a positivity bias in the later overall measure and subsequently exhibit increased purchase and loyalty behaviors toward the firm, as we found. However, if the overall judgment is elicited first, respondents tend to show a negativity bias in answering both the overall and the later specific questions (Schul and Schiff 1993). In this case, one may expect the mere-measurement effects to be mitigated for satisfied customers and accentuated for dissatisfied ones. In the present research, the same order—specific evaluations followed by the overall evaluation—was adopted for all experimental group participants. As a result, the order effects of measurement could not be teased apart. Future research should examine the role played by order of judgment elicitation more closely, perhaps through manipulation of order in which measures are presented to participants in a field or laboratory setting.

In this article, we examined the effect of responding to satisfaction surveys on behavior and found positive effects for the specific customer base. In contrast, Ofir and Simonson (2001) examined the effect of anticipating responding to satisfaction surveys and found negative effects on satisfaction, intentions, and behavior. They find this occurs because when consumers know they will evaluate their experience later, they focus on negative aspects of the experience. This suggests that the positive effects observed in our experiment might be reduced in situations in which consumers expect to have their satisfaction measured.

These results have implications for the conduct of applied marketing research, showing that employing marketing research has consequences that extend beyond the elicitation of customer perceptions and cause significant and sustained changes in customer behavior. At the first blush, these results appear to bode well for marketing research practice, demonstrating its unintended positive influence on customer behaviors. However, any enthusiasm stemming from such an interpretation must be tempered for three reasons. First, the positive consequences of measuring satisfaction are predicated on the valence of satisfaction expressed by the customer. The pattern of results found here reflects a highly satisfied customer base. It is possible that conducting such a study with a less satisfied customer base may result in a reversal of results, with negative effects on participating customers. In such cases, conducting marketing research may have deleterious consequences for the firm.

Second, these results are troubling from an interpretation standpoint: Can we rely on measures obtained from such studies when the act of measurement has so significantly changed the underlying phenomenon? In a related sense, these results suggest that even when the research is conducted using proper design and implementation procedures, care needs to be taken when extrapolating results from the surveyed group onto the general consumer population. Such effects may also affect longitudinal data collection or analysis as well, since participants are likely to be quite different from nonparticipants on account of their participation.

Third, our results raise some intriguing questions regarding ethics of applied marketing research. While considerable research has been conducted to examine this subject (see Giacobbe and Segal [2000] for a recent review), much of this work has focused on advertent transgressions. The research presented here shows that measurement-induced judgments may modify customer behaviors even when the research study is conducted rigorously and adheres to the marketing research code of conduct. In the present case, the customer satisfaction survey promoted more sales and loyalty for the firm, and such effects persisted long after the survey was over. It can be argued that this does not constitute sugging, since there was no volitional intent to sell (or even influence) on the researchers' part, nor did the instrument incorporate covert (or overt) sales messages. However, these findings do raise important questions regarding the unintended consequences of conducting marketing research. Devising a satisfactory strategy to mitigate such effects may involve aspects of instrument design (including a comprehensive set rather than some subset of specific questions, for instance), survey research methodologies, and instigation of correction processes during or after survey completion (through specific instructions and directions providing a description of this effect, for instance), all of which merit future thought and attention from both academics and practitioners of marketing research.

[*Received March 2001. Revised December 2001. David Glen Mick served as editor and Frank R. Kardes served as associate editor for this article.*]

REFERENCES

Allison, Paul D. (1995), *Survival Analysis Using the SAS System: A Practical Guide*, Cary, NC: SAS Institute.

Bagozzi, Richard P. and Utpal M. Dholakia (1999), "Goal Setting and Goal Striving in Consumer Behavior," *Journal of Marketing*, 63 (Special Issue), 19–32.

Bickart, Barbara A. (1993), "Carryover and Backfire Effects in Marketing Research," *Journal of Marketing Research*, 30 (February), 52–62.

Bowers, Diane K. (1995), "Sugging Banned, At Last," *Marketing Research*, 7 (4), 40.

Eagly, Alice H. and Shelly Chaiken (1993), *The Psychology of Attitudes*, Orlando, FL: Harcourt Brace.

Feldman, Jack M. and John G. Lynch, Jr. (1988), "Self-Generated Validity and Other Effects of Measurement on Belief, Attitude, Intention, and Behavior," *Journal of Applied Psychology*, 73 (3), 421–435.

Fitzsimons, Gavan J. and Vicki G. Morwitz (1996), "The Effect of Measuring Intent on Brand-Level Purchase Behavior," *Journal of Consumer Research*, 23 (June), 1–11.

Fitzsimons, Gavan J. and Baba Shiv (2001), "Nonconscious and Contaminative Effects of Hypothetical Questions on Subsequent Decision Making," *Journal of Consumer Research*, 28 (September), 224–228.

Fitzsimons, Gavan J. and Patti Williams (2000), "Asking Questions Can Change Choice Behavior: Does It Do So Automatically or Effortfully?" *Journal of Experimental Psychology: Applied*, 6 (3), 195–206.

Fornell, Claes (1992), "A National Customer Satisfaction Barometer: The Swedish Experience," *Journal of Marketing*, 56 (January), 6–21.

Giacobbe, Ralph W. and Madhav N. Segal (2000), "A Comparative Analysis of Ethical Perceptions in Marketing Research," *Journal of Business Ethics*, 27 (3), 229–245.

Gillund, Gary and Richard M. Shiffrin (1984), "A Retrieval Model for Both Recognition and Recall," *Psychological Review*, 91, 1–67.

Hastie, Reid (1984), "Causes and Effects of Causal Attribution," *Journal of Personality and Social Psychology*, 46 (1), 44–56.

Kardes, Frank R. (1988), "Spontaneous Inference Processes in Advertising: The Effects of Conclusion Omission and Involvement on Persuasion," *Journal of Consumer Research*, 15 (September), 225–233.

Kardes, Frank R., Gurumurthy Kalyanaram, Murali Chandrashekaran, and Ronald J. Dornoff (1993), "Brand Retrieval, Consideration Set Composition, Consumer Choice, and the Pioneering Advantage," *Journal of Consumer Research*, 20 (June), 62–75.

Maris, Eric (1998), "Covariance Adjustment versus Gain Scores—Revisited," *Psychological Methods*, 3, 309–327.

Morwitz, Vicki G. and Gavan Fitzsimons (2000), "The Mere-Measurement Effect: Why Does Measuring Purchase Intentions Change Actual Purchase Behavior?" working paper, New York University, New York, NY 10012.

Morwitz, Vicki G., Eric Johnson, and David Schmittlein (1993), "Does Measuring Intent Change Behavior?" *Journal of Consumer Research*, 20 (June), 46–61.

Ofir, Chezy and Itamar Simonson (2001), "In Search of Negative Customer Feedback: The Effect of Expecting to Evaluate on Satisfaction Evaluations," *Journal of Marketing Research*, 38 (May), 170–182.

Reichheld, Frederick F. (1996), *The Loyalty Effect*, Boston: Harvard Business School Press.

Reinartz, Werner J. and V. Kumar (2000), "On the Profitability of Long-Life Customers in a Non-contractual Setting: An Empirical Investigation and Implications for Marketing," *Journal of Marketing*, 64 (October), 17–35.

Ross, Lee and Richard E. Nisbett (1991), *The Person and the Situation: Perspectives of Social Psychology*, Philadelphia: Temple University Press.

Schul, Yaacov and Miriam Schiff (1993), "Measuring Satisfaction with Organizations: Predictions from Information Accessibility," *Public Opinion Quarterly*, 57 (4), 536–551.

Schuman, Howard and Stanley Presser (1980), "Public Opinion and Public Ignorance: The Fine Line between Attitudes and Nonattitudes," *American Journal of Sociology*, 85 (March), 1214–1225.

Sherman, Steven J. (1980), "On the Self-Erasing Nature of Errors in Prediction," *Journal of Personality and Social Psychology*, 39 (August), 211–221.

Sheth, Jagdish N. and Atul Parvatiyar (1995), "Relationship in Consumer Markets: Antecedents and Consequences," *Journal of the Academy of Marketing Science*, 23 (4), 255–271.

Simmons, Carolyn J., Barbara A. Bickart, and John G. Lynch, Jr. (1993), "Capturing and Creating Public Opinion in Survey Research," *Journal of Consumer Research*, 20 (September), 316–329.

Spangenberg, Eric R. and Anthony G. Greenwald (1999), "Social Influence by Requesting Self-Prophecy," *Journal of Consumer Psychology*, 8 (1), 61–89.

Tourangeau, Roger and Kenneth A. Rasinski (1988), "Cognitive Processes Underlying Context Effects in Attitude Measurement," *Psychological Bulletin*, 103 (3), 299–314.

Weiner, Bernard (1985), "'Spontaneous' Causal Thinking," *Psychological Bulletin*, 97 (1), 74–87.