MARKETING SCIENCE
Vol. 33, No. 2, March–April 2014, pp. 165–187
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

http://dx.doi.org/10.1287/mksc.2013.0815

Do Bonuses Enhance Sales Productivity? A Dynamic Structural Analysis of Bonus-Based Compensation Plans

Doug J. Chung
Harvard Business School, Boston, Massachusetts 02163, dchung@hbs.edu

Thomas Steenburgh
Darden Graduate School of Business, University of Virginia, Charlottesville, Virginia 22906, steenburght@darden.virginia.edu

K. Sudhir
Yale School of Management, New Haven, Connecticut 06520, k.sudhir@yale.edu

We estimate a dynamic structural model of sales force response to a bonus-based compensation plan. This paper provides substantive insight into how different elements of the compensation plan enhance productivity. We find evidence that (1) bonuses enhance productivity across all segments; (2) overachievement commissions help sustain the high productivity of the best performers, even after attaining quotas; and (3) quarterly bonuses help improve performance of the weak performers by serving as pacers to keep the sales force on track in achieving its annual sales quotas. The paper also introduces two main methodological innovations to the marketing literature: First, we implement empirically the method proposed by Arcidiacono and Miller [Arcidiacono P, Miller RA (2011) Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. *Econometrica* 79(6):1823–1867] to accommodate unobserved latent-class heterogeneity using a computationally light two-step estimator. Second, we illustrate how discount factors can be *estimated* in a dynamic structural model using field data through a combination of (1) an exclusion restriction separating current and future payoff and (2) a finite-horizon model in which there is no forward-looking behavior in the last period.

Keywords: sales force compensation; bonuses; quotas; dynamic structural models; two-step estimation; discount factors
History: Received: July 24, 2011; accepted: August 4, 2013; Preyas Desai served as the editor-in-chief and Miguel Villas-Boas served as associate editor for this article. Published online in *Articles in Advance* November 7, 2013.

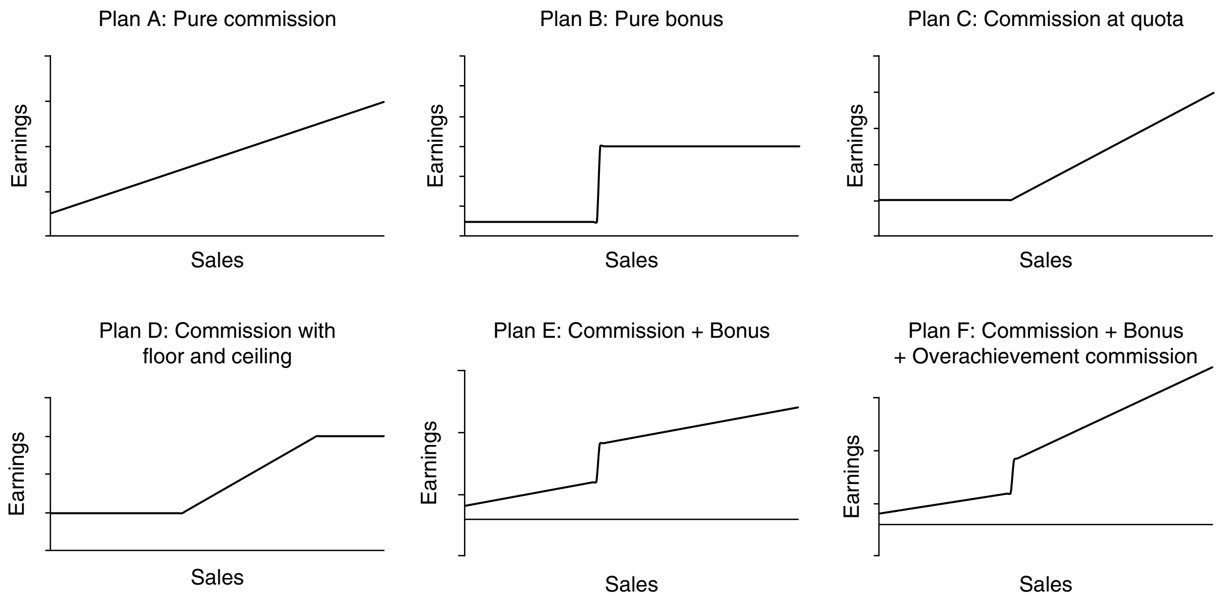
1. Introduction

Personal selling is one of the most important elements of the marketing mix, especially in the context of business-to-business (B2B) firms. An estimated 20 million people work as salespeople in the United States (Zoltners et al. 2008). Sales force costs average about 10% of sales revenues and as much as 40% of sales revenues for certain industries (Heide 1999). In the aggregate, U.S. firms spent more than $800 billion on sales forces in 2006, a sum three times larger than advertising spending (Zoltners et al. 2008).

Marketing researchers routinely create response models for marketing mix instruments such as price, sales promotion, and advertising. Meta-analyses of various research studies estimate that the sales force expenditure elasticity is about 0.34 (Albers et al. 2010), relative to about 0.22 for advertising (Assmus et al. 1984) and about −2.62 for price (Bijmolt et al. 2005). Relative sales force expenditure elasticity is useful in determining the relative effectiveness of different instruments in the marketing mix, but little insight is given on how to design a sales force compensation plan, which is widely understood to be the primary tool by which firms can induce their sales forces to exert the optimal levels of effort and thus to optimize the use of sales force expenditures.

A compensation plan can consist of many components: salary, commissions, and bonuses on achieving a certain threshold of performance called quotas. Figure 1 shows a variety of compensation plans that include combinations of these components. Which of these different types of contracts should a particular firm offer to its sales force to maximize profits? What combination of salary, commission, and/or quota-based bonuses should one use? Having chosen the compensation components, what should be the

Figure 1 Types of Incentive Compensation Schemes



specific parameters for commission rate, quotas, and bonus levels? Furthermore, what is the right frequency for quota targets? For example, should there be quarterly or annual quotas? A firm needs to understand how the sales force will respond to different elements to develop an appropriate compensation plan.

Thus, this paper has two substantive goals: First, it seeks to gain insight into how a firm should design its compensation plan. For example, should a firm offer quotas and bonuses in addition to commissions? Second, we aim to determine how often quotas should be set and bonuses paid. For example, should a firm implement a monthly, quarterly, or annual bonus? Should it offer a quarterly bonus in addition to an annual bonus? In the education literature, researchers have argued that frequent testing leads to better performance outcomes (Bangert-Drowns et al. 1991). Can quarterly quotas serve a similar role to improve outcomes? As in education, where frequent exams help students to be better prepared for the comprehensive final exam, frequent quota-bonus plans may serve as a mechanism to keep the sales force motivated to perform well in the short run so as to be within striking distance of the overall annual performance quota.

Quotas and bonuses are widely used by firms. According to Joseph and Kalwani (1998), only about 24% of firms use a pure commission-based plan; the rest used some form of quotas. As per the 2008 Incentive Practices Research Study by ZS Associates, 73%, 85%, and 89% of firms in the pharma/biotech, medical devices, and high-tech industries, respectively, use quota-based compensation (*Training* 2008). Yet, despite the ubiquity of quota-based compensation, there is considerable controversy in both the theoretical and empirical literatures about the effectiveness of quotas and bonuses relative to straight linear commission plans.

We begin with a discussion of the relevant theoretical literature. Using the principal-agent framework of Holmstrom (1979), Basu et al. (1985) and Rao (1990) find that the combination of salary and commission (usually nonlinear with respect to sales) is optimal. In this light, quota-bonus plans can be seen as an approximation to a continuous nonlinear plan that also takes into account heterogeneity in territory sales potential. However, under the assumption of linear exponential utility and normal errors, Holmstrom and Milgrom (1987) and Lal and Srinivasan (1993) show that a linear commission scheme can achieve the first-best outcome. Yet why do we see quota-bonus plans? Raju and Srinivasan (1996) suggest that even though a commission-over-quota plan may not be theoretically optimal, it provides the best compromise between efficiency and ease of implementation. Others argue that quota-based plans offer high-powered incentives that can motivate salespeople to work harder (e.g., Darmon 1997). Park (1995) and Kim (1997) demonstrate that a quota-bonus plan may lead to the first-best outcome, but in their framework, a quota-bonus plan is just one of many possible plans that can lead to first-best outcomes. Oyer (2000) shows that when participation constraints are not binding, a quota-bonus plan with linear commissions beyond quotas can be uniquely optimal because it can concentrate the compensation in the region of effort where the marginal revenue from effort minus the cost of compensation is maximized.

In terms of empirical work, Ferrall and Shearer (1999) and Paarsch and Shearer (2000) estimate static structural models of worker behavior given linear contracts. Oyer (1998) is the first to empirically investigate quota-based plans. Using aggregate sales across different industries in different quarters, he concludes that quota-based plans encourage salespeople to maneuver the timing of orders; this negative effect overwhelms the positive benefit of quotas. Using reduced-form analysis of individual-level data, Steenburgh (2008) finds that the net improvement in revenues from effort dominates the inefficiencies induced by intertemporal dynamic considerations and shows that an aggregate analysis might have led to the opposite conclusion similar to that of Oyer (1998).

Copeland and Monnet (2009) estimate the first dynamic structural model of worker productivity in a check-sorting environment with nonlinear incentives. Unlike sales force productivity, where we observe only aggregate sales, the outcomes associated with every processed check are observed in their environment; hence, they have a qualitatively different model. Our paper shares many similarities with a recent paper by Misra and Nair (2011), who also estimate a dynamic structural model of sales force compensation, although they use a very different quota-compensation scheme. In contrast to our focus on quotas with bonuses (plan F in Figure 1), Misra and Nair analyze quotas with floors and ceilings on commissions (plan D in Figure 1). They conclude that quotas reduce performance. This is because of two characteristics of their quotas: First, the quota ceiling (beyond which salespeople receive zero additional compensation) limits the effort of the most productive salespeople, who would normally have exceeded that ceiling. Second, the company followed an explicit policy of ratcheting quotas based on past productivity. This reduces salespeoples' incentives to work hard in any given period, because hard work is penalized through higher future quotas. In contrast, we find that quotas coupled with bonuses enhance performance. In the plan we consider, the company offers extra overachievement commissions for exceeding quotas and uses a group quota updating procedure that minimizes ratcheting effects. Thus, these two papers offer complementary perspectives that enhance our understanding of how quotas impact performance.

From a methodological perspective, our paper introduces two key ideas to the marketing literature. First, we accommodate latent-class heterogeneity within the two-step conditional choice probability (CCP) estimation framework—an issue that has been an econometric challenge for the literature for nearly two decades. Misra and Nair (2011) avoid the unobserved heterogeneity issue by estimating each salesperson's utility function separately.[1] Although the use of two-step estimation approaches has recently gained popularity (Hotz and Miller 1993, Bajari et al. 2007) because of their ease of computation relative to traditional nested fixed-point estimation approaches (e.g., Rust 1987), their use in empirical applications has been limited by their inability to accommodate unobserved heterogeneity. Arcidiacono and Miller (2011) propose an algorithm that accommodates latent-class heterogeneity within the two-step framework. Our paper introduces this idea to the marketing literature and illustrates its implementation with an empirical application that accommodates unobserved heterogeneity using two-step dynamic structural estimation.[2]

Second, unlike Misra and Nair (2011)—and of broader importance to the dynamic structural modeling literature—we *estimate* rather than *assume* discount factors. It is well known in the literature on dynamic structural models that discount factors cannot be identified in standard applications because there are no instruments that provide exclusion restrictions across current and future period payoffs (Rust 1994). Hence the standard approach is to assume discount factors. We illustrate how the bonus setting provides us exclusion restrictions allowing us to estimate discount factors—thus demonstrating how field data can indeed be used to estimate discount factors.[3] Since bonus payoffs occur only at the end of each quarter or year, in the nonbonus periods, the probability of achieving the quota and receiving a bonus will not affect the current payoff, only future payoffs. Only a forward-looking person (i.e., one with a nonzero discount factor) would be concerned with whether she is close to the quota in nonbonus periods. We demonstrate through reduced-form evidence that such behavior exists in the data, and then we exploit this exclusion restriction to identify the discount factor. In addition, the finite-horizon nature of the sales force problem implies that, in the last period, the model reduces to a static model, for which it is well known that utility can be identified. This further facilitates identification of the discount factor. Yao et al. (2012) apply our idea to a cellphone usage context to identify discount factors. In their setting, how

[1] This is similar to estimating individual-level utility functions in conjoint analysis or scanner panel data, when there are a large number of observations per individual. More importantly, the approach requires that salespeople exert effort equally across all customers—an assumption they show is valid in their data but unlikely to hold in general. Our latent-class approach works in the more common situation where there are limited observations per individual.

[2] Finger (2012) and Beauchamp (2012) are two concurrent working papers implementing Arcidiacono and Miller's approach in economics.

[3] Use of exclusion restrictions to empirically infer discount factors is important because, prior to this paper, the conventional wisdom was that field data cannot be used for discount factor identification and that one needs to use surveys (e.g., conjoint data as in Dubé et al. 2012).

close the user is to exceeding his or her monthly quota serves as an exclusion restriction, much like in our setting. Similar to the last period of our data, where the model reduces to a static model, they have data for a period in which pricing is linear and therefore reduces to a static model.

There are three specific modeling and estimation challenges in the structural estimation of response to compensation plans, especially those with quotas and bonuses. First, we do not observe the effort of the sales force, only the outcome of the agent's effort (i.e., sales), which is correlated with effort. This requires a modeling assumption on the link between sales and effort. The identification of effort poses particular challenges in our application as a result of seasonality and potential sales substitution across quarters.[4] We discuss these issues in §2.3. Second, compensation plans do not change over time. Here, we draw on an empirical insight from Steenburgh (2008) for identification. In any given period, a sales agent's optimal effort depends on her state—how close the person is to achieving her quota. A sales agent may find it optimal to reduce her effort when she is close or very far from achieving the quota, but she may stretch herself to reach the quota when she has a moderate chance of achieving it. Thus changes in sales in response to changes in the agent's state (*distance to quota*) within and across agents facilitate identification. Finally, quotas and bonuses induce dynamic forward-looking behavior; an agent choosing to exert effort has to be concerned not just with the current payoff but with the effect of that effort on the likelihood of obtaining a bonus in the future. This requires a dynamic structural model.

We estimate a dynamic structural model of sales force response to various features of the compensation plan using sales force output and compensation data from a Fortune 500 firm that sells office durable goods. This firm uses plan F in Figure 1. In addition, bonuses are provided at two different frequencies: quarterly and annual. As the compensation structure of the focal firm features almost all dimensions in typically used compensation plans, we observe how the sales force responds to the plan's different dimensions. This "rich" plan provides us with two key benefits: First, the presence of bonuses helps us identify and estimate discount factors. Second, even though in theory one can perform counterfactuals of any type of compensation plan if able to estimate structural parameters (other than discount factors) for a salesperson with a less rich compensation plan, an analyst or manager should have greater faith in the counterfactuals based on parameters that were estimated from observed responses to different elements of the compensation plan.

The rest of the paper is organized as follows. Section 2 discusses the institutional details of the compensation plan at the firm, provides a numerical example that offers insights into how bonuses induce effort, and provides some model-free evidence that facilitates model building. We present the model and the estimation methodology in §§3 and 4, respectively. Section 5 discusses the estimation results and the counterfactual analysis. Section 6 concludes.

2. Institutional Details and Descriptive Analysis

We first describe the details of the bonus-based compensation plan, followed by a numerical example to clarify how bonuses can help motivate the salesperson and induce intertemporal effort that assists us with identification. We then provide model-free evidence of forward-looking behavior, seasonality, and the absence of sales substitution across quarters.

2.1. The Compensation Plan

The focal firm under study is a highly regarded multinational Fortune 500 company that sells durable office products primarily using its own direct sales force. Each sales agent is given an "exclusive" territory; the firm traditionally does not encourage group work or team cooperation among the sales force. The firm also has an indirect sales force through "rep" firms who do not compete with the direct sales force. They are paid purely on commission, unlike the regular sales force.[5]

Our analysis focuses on sales performance data from 348 salespeople from the regular sales force during the three-year period 1999–2001. The firm's compensation structure follows the pattern in plan F of Figure 1, and the details of the compensation schedule for the period of analysis are described in Table 1. We provide descriptive statistics of the data in Table 2.

Every month, salespeople receive a fixed monthly salary (on average, $3,585) and a commission of 1.5% of the revenues generated in that month. In the first three quarters, a quarterly lump-sum bonus of $1,500 is paid if the quarterly quotas are met. At the end of the year (i.e., end of the fourth quarter), an annual lump-sum bonus of $4,000 is paid if the annual quota is met. Furthermore, an overachievement commission of 3% is paid for any excess revenues beyond the annual quota. There are no caps on revenues for which an agent could obtain commissions or overachievement commissions. Overall, for a salesperson that meets all quotas, the salary component will be roughly 50% of total compensation.

[4] This issue has parallels in empirical channel response models. For example, Sudhir (2001) makes an inference about manufacturer actions (wholesale prices) from the observed retail price and sales to infer competition between manufacturers.

[5] Such rep firms are the focus of Jiang and Palmatier (2009).

Table 1 Firm's Compensation Plan

Type	Description	Payment period
Quarterly bonus	\$1,500 awarded if the quarterly revenue exceeds quarterly quota	March, June, September
Annual bonus	\$4,000 awarded if the annual revenue exceeds annual quota	December
Base commission	About 1.5%[a] paid in proportion to the revenue generated each month	Every month
Overachievement commission	About 3%[a] paid in proportion to the total cumulative revenue surpassing the annual quota	December

[a]These numbers are approximate for confidentiality reasons.

In building annual and quarterly quotas for its sales force, the company uses internal metrics called "monthly allocated quotas" (based on expected monthly revenues, given seasonality, and territorial characteristics), though these are not used for compensation. We do not use these quotas for our modeling and estimation, but we use them to benchmark performance in the reduced-form analysis.

The most important element in performance evaluation within the firm is the annual quota; i.e., the firm views a salesperson as having a successful year if the annual quota is met. From Table 2, we see that salespeople meet their annual quota roughly 50% of the time. All quotas, including the quarterly quotas, are updated annually and reset in January of the following year. However, managers at the firm informed us that they were sensitive to the fact that ratcheting quotas based on individual performance could lead the sales force to purposefully reduce performance. To avoid such an adverse impact on sales, the quota adjustment year after year was done based only on group performance, where each individual's current performance would have minimal direct impact on her future quota. Nevertheless, we test for statistical evidence of ratcheting in the data. Table 3 reports the results of the tests. In Model 1, we regress the percentage increase in the annual quota on whether the salesperson met her quota in the previous year. We find no significant effect of meeting the previous year's quota on the current year's quota. In Model 2, we regress the logarithm of the annual quota on whether a salesperson met her quota in the previous year, this time controlling for salesperson fixed effects. Again, we find no significant effect. Finally, in Model 3, we regress the logarithm of quota on the logarithm of the previous year's performance relative to the annual quota. Again, we find no significant effects. Thus, consistent with the intuition of the firm's managers, who sought to avoid the negative impact of ratcheting on employee performance, we find no direct statistical evidence of ratcheting. For this application, we therefore abstract away from quota ratcheting and treat each salesperson's quarterly and annual quotas as exogenous. To the extent that a salesperson's quota is partly updated based on her own performance, there are still potential ratcheting effects that may be detectible with longer panel data.[6]

Table 2 Descriptive Statistics

No. of salespeople		348	
Average salary (US\$)		3,585	
Average tenure (years)		11.8	
	Average quota (US\$ '000)	% achieving quota	Average sales (US\$ '000)
Quarter 1	232.4	51.1	273.0
Quarter 2	374.2	49.8	445.0
Quarter 3	397.1	42.8	407.0
Quarter 4	—	—	565.6
Annual	1,639.3	49.9	1,690.6

Table 3 Testing for Ratcheting Effects

	Model 1	Model 2	Model 3
Dependent variable	$(AQ_y - AQ_{y-1})/AQ_{y-1}$	$\log(AQ_y)$	$\log(AQ_y)$
$I(Sales_{y-1} \geq AQ_{y-1})$	0.0364 (0.0229)	0.0033 (0.0299)	
$\log(\%AQ_{y-1})$			0.0347 (0.0196)
Year fixed effects	Yes	Yes	Yes
Salesperson fixed effects	NA	Yes	Yes

Notes. Standard errors are shown in parentheses. y represents each year in this analysis.

2.2. Numerical Example

Before building an empirical model of salesperson's response to a bonus-based compensation scheme, we provide a numerical example that illustrates how a quota-bonus scheme can be more effective than pure commissions in generating more effort. The example will also show that a person's distance to quota

[6] The assumption of no ratcheting allows us to make the finite-horizon argument for identification of the discount factor discussed in §4.3. The exclusion restriction argument for identification will be valid even if we allow for ratcheting.

can induce variations in optimal effort (and generated sales)—providing intuition for our identification strategy.

Let the utility function of the salesperson that trades off effort (e) and income from sales (s) and who has sold S units at the beginning of the new period be

$$U(s, e, S) = -de^2 + rs + BI_{\{s+S\geq Q\}},$$

where $-d$ is the disutility parameter and r is the commission rate ($d > 0, r > 0$), and B is the bonus for reaching quota (Q). For illustration, let us assume a direct match between sales and effort; i.e., $s = e$.

First, we illustrate the effectiveness of bonuses using a static model. For simplicity, assume $S = 0$. In the pure commission case with no bonus, where $d = 1$, $r = 10$, and $B = 0$, the optimal effort is $e^* = 5$. In the bonus case, where $Q = 10$ and $B = 30$, the optimal effort is higher, at $e^* = 10$, and the compensation cost to the firm is \$130. To achieve the same level of sales and effort ($e^* = 10$) from a pure commission plan, the commission rate r has to increase to 20 and costs more for the firm, at \$200. Figure 2(a) illustrates these results graphically. Thus, the quota-bonus plan in this context induces more effort and sales for the same level of compensation. We are not claiming that a bonus-based plan is the optimal compensation plan. Through the example above, we are simply illustrating that a nonlinear plan can be more efficient than a linear plan, suggesting that the optimal compensation plan is possibly nonlinear and that the bonus plan can be a way to approximate the optimal nonlinear incentive plan.

Second, we show how distance to quota induces variation in effort. Let $d = 2$, $r = 10$, $Q = 10$, and $B = 30$. Consider three scenarios of distance to quotas: $S = 0$ (far away from quota), $S = 5$ (moderately close to quota), and $S = 7$ (close to quota). Figure 2(b) shows that the optimal effort levels are $e^* = 2.5$, 5, and 3, respectively, for the three cases; the salesperson exerts maximum effort at $S = 5$, when moderately far away from quota, all else being equal. We use this variation of optimal levels of effort (thus, sales) within agents across time to identify their preferences in our main model.

2.3. Model-Free Analysis

We consider three features of the data that inform model development. First, we provide evidence of forward-looking behavior induced by bonuses and hence the need to develop a dynamic model. Second, we demonstrate the presence of seasonality in the data, which therefore requires us to accommodate this in the model. Third, we show that sales substitution across quarters by sales agents appears to be limited in our data; we therefore abstract away from modeling the timing of sales bookings by sales agents.

2.3.1. Forward-Looking Behavior. As discussed previously, bonuses provide an exclusion restriction in that they do not impact current payoffs, only future payoffs. To the extent that a sales agent's performance is affected by variables relating to proximity

Figure 2(a) **How Quotas and Bonus Serve as Stretch Goals**

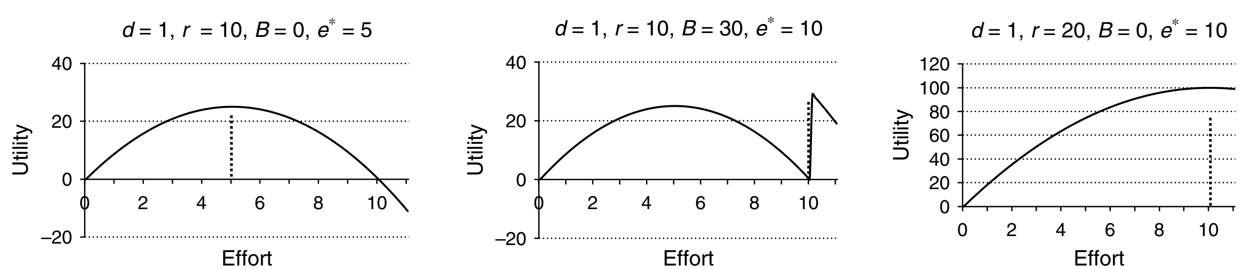


Figure 2(b) **Effort as a Function of Distance to Quotas**

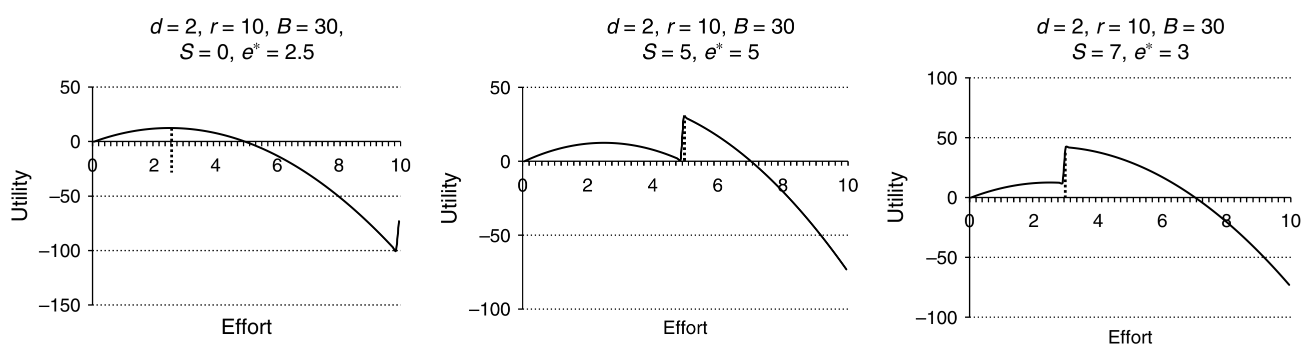
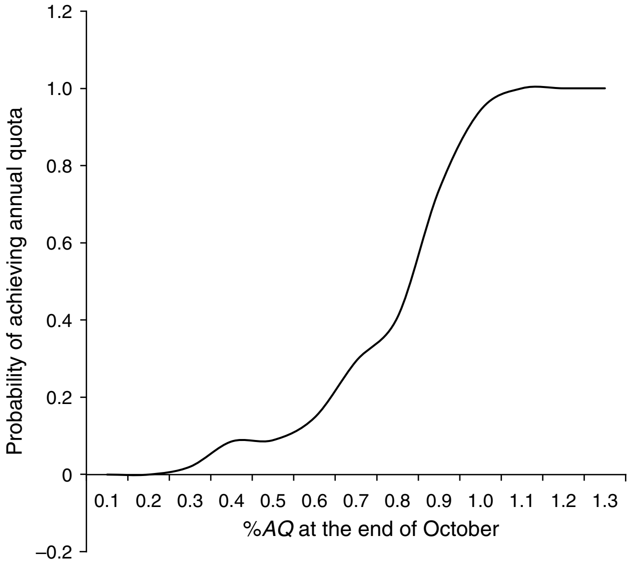


Figure 3 Fraction of People Achieving Quota in December



to bonuses, this is evidence of forward-looking behavior.[7] But proximity to bonus quota will impact performance only if agents have a reasonable chance of making the quota. Figure 3 shows the graph of the probability of reaching the annual quota, conditional on the cumulative fraction of annual quota (%AQ) achieved by November. It is clear that there is very little chance of achieving quota if %$AQ < 0.5$. We therefore divide agents by their state %$AQ < 0.5$ and %$AQ > 0.5$ to test whether the state affects sales and estimate regressions on sales performance in November as a function of their states. Table 4 reports the results of the regressions. Consistent with forward-looking behavior, the state %AQ is significant only for agents with %$AQ > 0.5$.[8]

For additional evidence of forward-looking behavior, we show scatterplots and the best-fitting nonparametric smoothed polynomial (and its 95% confidence interval) of sales revenues normalized by monthly allocated quotas in the quarterly bonus months (March, June, September, and December) against the percentage of quota attained by the previous month in Figure 4(a). For March, June, and September, the x axis is the percentage of quarterly quota completed (%QQ), whereas for December, the x axis is the percentage of annual quota completed (%AQ). The vertical line shows the %QQ and %AQ at which salespeople, on average, achieve their monthly allocated quotas.

[7] A similar argument is made in providing evidence of forward-looking behavior by students in the textbook market by Chevalier and Goolsbee (2009).

[8] The qualitative conclusions are robust in the range of thresholds of %AQ from 0.4 to 0.6. Note that if there were no overachievement commission, agents very close to the quota may reduce their effort.

Table 4 Sales Performance in November

	Estimate	
	%$AQ < 0.5$	%$AQ > 0.5$
Intercept	0.05*** (0.018)	0.06*** (0.0123)
%AQ	0.06 (0.049)	0.04*** (−0.0126)

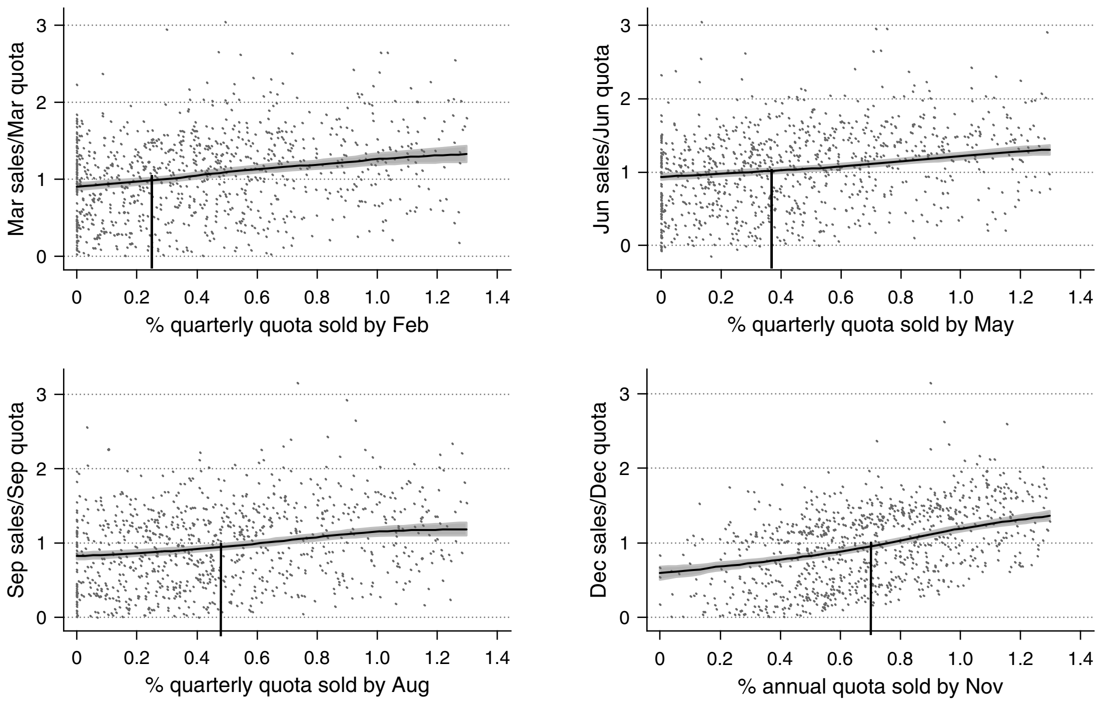
Note. Standard deviations are shown in parentheses.
***$p < 0.01$.

Two key elements stand out from Figure 4(a). First, across the board there is little reduction in effort when salespeople are close to achieving quota, most likely as a result of the overachievement commission rate. Second, there is a steady increase over time in the %QQ and %AQ threshold beyond which salespeople reach their monthly targets. The threshold is about 25% in March, 35% in June, 45% in September, and close to 70% in December. Early in the year, even if below targets, salespeople still have hopes of receiving a large annual bonus by working hard. As one gets closer to year-end, the chances of reaching the quota become less likely, and salespeople respond by reducing effort even at higher levels of %AQ and %QQ. Since annual bonuses should have no impact on current payoffs in March, June, or September, only on future payoffs, this is suggestive of forward-looking behavior.

The next set of graphs presented in Figure 4(b) shows the same relationship in the pre-bonus months (February, May, August, and November) and provides additional evidence for forward-looking behavior. In February, May, and even August, at all levels of %QQ, the salesperson on average sells above the monthly allocated quota. This is because hard work (and some good luck in the form of positive sales shocks) may give the salesperson a reasonable chance of attaining the smaller quarterly targets. However, in November, only at a very high level of %AQ does the salesperson sell above the monthly allocated quota, because one has a very limited chance of closing the large gap in just two months. In the pre-quarterly bonus months, the immediate future quarterly bonus impacts behavior, even though it has no impact on the current payoff; this indicates more conclusive forward-looking behavior.

This evidence leads to a natural question: Should the large annual bonus be split into a quarterly bonus (as in other months) *and* an annual bonus? The quarterly bonus can prevent salespeople from giving up in November, even if they do not have a chance of reaching the annual quota. But with such a quarterly quota, early in the year, agents may have limited incentive to work hard after reaching quarterly quotas. How these

Figure 4(a) Sales and Percentage Quota Achieved—Bonus Months



two issues trade off is an empirical question that we subsequently address in the counterfactual analysis.

2.3.2. Seasonality. Figures 5(a) graphs the average revenues over the months for the regular sales force. There are clear peaks at the end of each quarter. These peaks could be due to either seasonality or bonuses at the end of each quarter. Figure 5(b) shows the index of indirect sales revenue (ISR) for

Figure 4(b) Sales and Percentage Quota Achieved—Pre-Bonus Months

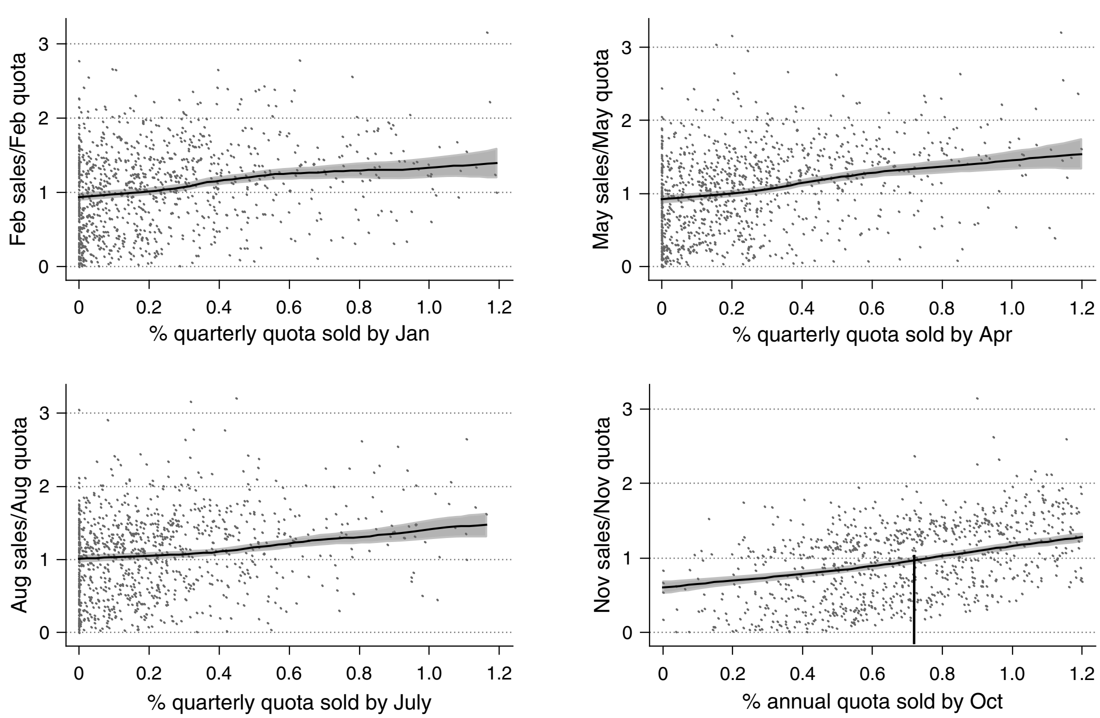


Figure 5(a) **Revenues from the Regular Sales Force**

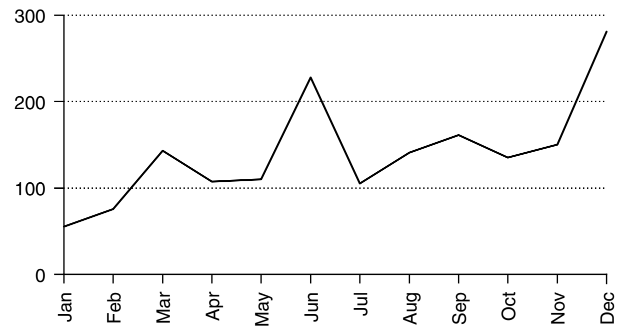
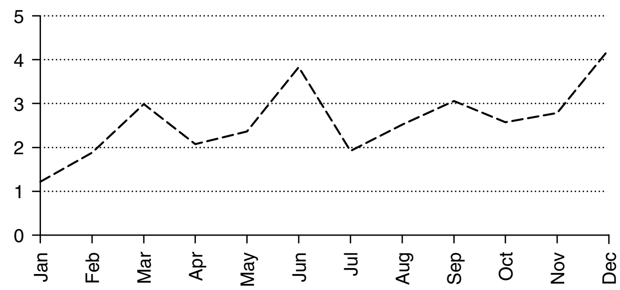
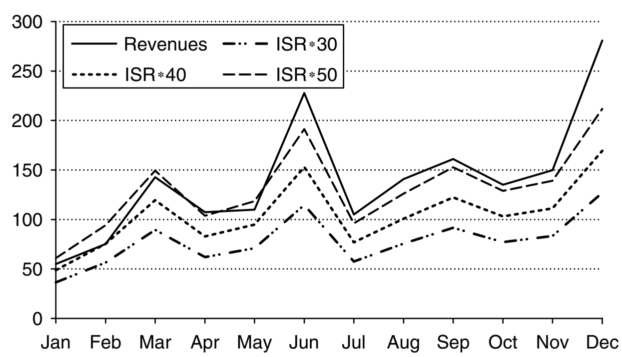


Figure 5(b) **ISR Index**



Notes. The ISR index is on a base of 1, reflecting the indirect sales in January 1999. Here, we have averaged the index across three years (1999–2001).

Figure 5(c) **Revenues from the Regular Sales Force and ISR Index Multiples**



the pure commission indirect sales force.[9] Because the ISR index is not contaminated by bonuses, we use it to control for seasonality and isolate the effect of bonuses on salesperson effort and revenues. To build intuition for how the ISR index can help control for seasonality and isolate effort, see Figure 5(c), which graphs the average revenues of the direct sales force and multiples of the ISR index. At a multiple of around 50, the ISR virtually mimics the average revenues, making the revenues from the commissioned and bonus sales force close to identical. This suggests that bonuses are not effective in inducing additional effort. When the ISR index has a multiple of 30 or 40, even after the overall seasonality is accounted for, there is a gap in revenue that we interpret as induced by effort. It is interesting that these gaps are larger at the end of the quarter, suggesting the value of bonuses in inducing effort. We empirically estimate the multiple for the ISR index to control for seasonality. We later show in §5.1 that the ISR multiple for our model is about 25, given that the average of lagged annual quota is 1,639. We acknowledge that accounts handled by the direct and indirect sales force are likely systematically different, but our maintained assumption is that the seasonality multipliers on sales are identical across the two types of accounts. We acknowledge that this assumption is a limitation to fully control for seasonality.

Table 5 **Testing for Sales Substitution Across Quarters**

	Model 1	Model 2
Other months of quarter	168.87***	101.76***
	(34.57)	(34.35)
First month of quarter	147.79***	101.86***
	(34.87)	(34.52)
Other months of quarter ×	91.09***	56.58***
Previous month % distance to quota	(5.62)	(5.90)
First month of quarter ×	2.59	0.15
Previous month % distance to quota	(7.02)	(6.93)
ISR (indirect revenue) index		38.72***
		(2.26)
Salesperson fixed effects	Yes	Yes

Note. Standard deviations are shown in parentheses.
***$p < 0.01$.

2.3.3. Sales Substitution Across Quarters. One possibility is that salespeople giving up at the end of the quarter may be doing so to increase the odds of meeting quotas in subsequent quarters by simply not booking the sales in the current quarter. Alternatively, a salesperson who is meeting the quarterly quota may simply shift sales to the next quarter to increase the odds of meeting next quarter's quota. In either case, we should see a positive relationship between the first month of the quarter and the previous quarter's percentage cumulative performance to quota (%QQ or %AQ). However, in Table 5, we find that the coefficient of *First month of quarter* × *Previous month (%) distance to quota* is statistically insignificant and very small in magnitude.[10] We therefore abstract away from the challenges of modeling sales substitution in our application. We acknowledge that our test may have limited power in identifying timing behavior; future research should explore more seriously the issue of sales timing and the attendant modeling challenges.

[9] For the indirect sales force, the firm only provided us with an index of revenues. The ISR index is set to a base of 1 for January 1999.

[10] Managers of the focal firm stated that they monitor sales agents extensively to prevent such time shifting of sales bookings. Also, customers play a major role in determining when sales occur in this environment.

3. Model

Based on the model-free evidence, we build a dynamic model of sales force response to the quota-based compensation scheme. The timing of the model is as follows:

1. At the beginning of each year, the firm chooses an annual compensation plan.
2. Each month, agents observe their current state and exert effort in a dynamically optimal manner.
3. An idiosyncratic sales shock is realized; the shock plus the agent's effort determines the agent's realized sales for the period. The agent receives compensation.
4. The realized sales of the current period affect the agent's state of the next period. Steps 2 and 3 are repeated each month until the end of the year, and Steps 1–3 are repeated each year.

We describe the model in five parts: (i) the compensation plan, (ii) the sales agent's utility function, (iii) state transitions, (iv) the effort as a function of state variables, and (v) the optimal effort choice by the sales agent.

3.1. Sales Response Model

We model the sales revenue function (S_{it}) for salesperson i at time t in two parts: (1) a base level of sales independent of effort, parameterized by demand shifters (z_{it}^D), and (2) sales induced as a result of effort (e_{it}), parameterized by effort shifters that include territory and salesperson characteristics (z_{it}^E):

$$S_{it} = f(z_{it}^D) + e_{it}(z_{it}^E) + \varepsilon_{it}, \tag{1}$$

where ε_{it} is an additive sales revenue shock not anticipated by the salesperson when choosing effort.

As discussed previously, the market potential varies across territories and across time. To account for the cross-sectional variation in market potential, we use the annual quota from the previous year ($AQ_{i,\,y-1}$). To account for seasonality of demand across months, we use the ISR index (ISR_t). We also include an interaction between the two variables to allow for seasonality to have a different impact on differently sized territories.

For effort shifters in e_{it}, we use the following variables: given that effort is a function of demand shifters, we include both $AQ_{i,\,y-1}$ and ISR_t in z_{it}^E. As discussed in the motivation, the salesperson's state with respect to achieving her quota will have an impact on the effort she expends. We therefore use the cumulative percentage of quarterly and annual quotas completed till time t ($\%QQ_{it}, \%AQ_{it}$) as variables that affect effort. In addition, we use the time (month type) within quarterly and annual quota periods to allow for different effort levels at different temporal distances to bonus payments; thus, the effort policy function is different for each month. Furthermore, we allow a time-invariant salesperson-specific variable, tenure with the firm (τ), to moderate the level of effort.

Note that unlike the demand-shifter function $f(\cdot)$, which is common across all salespeople, the effort function will vary across salespeople. Specifically, we allow for salespeople to belong to one of multiple discrete segments; hence these effort functions will be estimated at the segment level. We estimate the effort function nonparametrically using Chebyshev polynomials of the variables described above.

3.2. Compensation Plan

The compensation plan has three components: (i) the monthly salary w_{it}; (ii) the end-of quarter bonus B_{iqt} received for achieving the corresponding quarterly quota Q_{iqt}, and the end-of-year bonus B_{iyt} received for achieving the corresponding annual quota Q_{iyt}; and (iii) a commission rate, r_{it}, per dollar's worth of sales and an overachievement commission rate, r'_{it}, given at the end of the year for sales over and above the annual quota for each individual i at time t. We represent the compensation plan for a salesperson i by the vector $\psi_{it} = \{w_{it}, Q_{iqt}, Q_{iyt}, B_{iqt}, B_{iyt}, r_{it}, r'_{it}\}$.

3.3. Salesperson's Per-Period Utility

In each period t, salesperson i receives a positive utility of wealth W_{it} earned based on realized sales and a disutility $C(e_{it};\theta_i)$ from exerting effort e_{it}. Thus the utility function is defined as

$$\begin{aligned} &U(e_{it}, S_{it}; \psi_i, \theta_i, \gamma_i) \\ &\quad = E[W(S_{it};\psi_i)] - \gamma_i \operatorname{var}[W(S_{it};\psi_i)] - C(e_{it};\theta_i), \end{aligned}$$

where γ_i and θ_i are the risk aversion and disutility parameters, respectively, for salesperson i.[11]

Given the sales levels and the compensation plan, the wealth for individual i, W_{it}, can be computed; W_{it} arises from four components: the per-period salary component w_{it}, the lump-sum bonus component B_{it}, the commission component C_{it}, and the overachievement commission component OC_{it}. The detailed expressions of wealth is as follows:

$$\begin{aligned} W_{it} &= w_{it} + B_{it} + C_{it} + OC_{it}, \\ B_{it} &= I_{qt} I\left(z_{i1t} + \frac{s_{it}(e_{it}(z_{it}^E), z_{it}^D; \alpha_i) + \varepsilon_{it}}{Q_{iqt}} > 1\right) B_{qt} \\ &\quad + I_{yt} I\left(z_{i2t} + \frac{s_{it}(e_{it}(z_{it}^E), z_{it}^D; \alpha_i) + \varepsilon_{it}}{Q_{iyt}} > 1\right) B_{yt}, \end{aligned}$$

[11] In the case of the constant absolute risk-aversion utility function (exponential utility function) with normal errors and a linear compensation plan, this functional form represents the certainty equivalent utility of the agent. Here, we consider the utility function to be a second-order approximation to a general concave utility function with a constant level of risk aversion.

$$C_{it} = (s_{it}(e_{it}(z_{it}^E), z_{it}^D; \alpha_i) + \varepsilon_{it}) r_{it},$$
$$OC_{it} = I_{yt} I\left(z_{i2t} + \frac{s_{it}(e_{it}(z_{it}^E), z_{it}^D; \alpha_i) + \varepsilon_{it}}{Q_{iyt}} > 1\right) \cdot \left(z_{i2t} Q_{iyt} + s_{it}(e_{it}(z_{it}^E), z_{it}^D; \alpha_i) + \varepsilon_{it} - Q_{iyt}\right) r'_{it},$$

where z_{i1t} and z_{i2t} are the percentage of quarterly and annual quotas completed, respectively, by salesperson i until time t; I_{qt} and I_{yt} are indicators for whether time t is a quarterly or annual bonus period.

In our empirical analysis, we use a quadratic functional form for the disutility function; specifically, $C(e; \theta_i) = \theta_i e^2$. Thus the set of structural parameters of the salesperson's utility function that need to be estimated is $\mu_i = (\theta_i, \gamma_i)$.

3.4. State Variables

As discussed previously, the nonlinearity of the compensation scheme with quotas and bonuses introduces dynamics into the sales agent's behavior because there is an additional trade-off between the disutility of effort today and a higher probability of a lump-sum bonus and overachievement commissions tomorrow. To incorporate the dynamics of the model, we consider the following key state variables: the percentage of annual quota completed, the percentage of quarterly quota completed, month type within the quarterly quota period, and month type within the annual quota period (time of the year). These state variables evolve as follows:

1. Percentage of quarterly quota completed (%QQ)

$$z_{i1t} = \begin{cases} 0 & \text{if } t \text{ is the start of the quarterly quota period,} \\ z_{i1(t-1)} + \dfrac{S_{i(t-1)}}{Q_{iqt}} & \text{otherwise;} \end{cases}$$

2. Percentage of annual quota completed (%AQ)

$$z_{i2t} = \begin{cases} 0 & \text{if } t \text{ is the start of the annual quota period,} \\ z_{i2(t-1)} + \dfrac{S_{i(t-1)}}{Q_{iyt}} & \text{otherwise;} \end{cases}$$

3. Month type within quarterly quota period

$$z_{i3t} = \begin{cases} 1 & \text{if } t \text{ is the start of the quarterly quota period,} \\ z_{i3(t-1)} + 1 & \text{otherwise;} \end{cases}$$

4. Month type within annual quota period

$$z_{i4t} = \begin{cases} 1 & \text{if } t \text{ is the start of the annual quota period,} \\ z_{i4(t-1)} + 1 & \text{otherwise.} \end{cases}$$

Whereas the first two state variables evolve stochastically, conditional on the effort levels and revenue shocks of the previous periods, the latter two evolve in a purely deterministic manner. Other state variables would include time-varying demand shifters, the ISR index, and territory characteristics for which we use the previous year's annual quota. We use tenure with the focal firm (τ) as an individual state variable that impacts effort. These state variables are collected in a state vector,

$$z_{it}^E = \{z_{i1t}, z_{i2t}, z_{i3t}, z_{i4t}, IS_t, AQ_{i(y-1)}, \tau_i\}.$$

3.5. Optimal Choice of Effort

Given the parameters of the compensation scheme ψ, as well as the state variables and their transitions, each sales agent would choose an effort level conditional on her states to maximize the discounted stream of expected future utility flow. Alternatively, if this value function is below the reservation wage, the salesperson may choose to leave the firm.

The stream of utility flow, under the optimal effort policy function, conditional on staying at the firm, can be represented by a value function:

$$V(z; \psi, \Omega) = \max_e \Big\{ U(e, z; \psi, \mu) + \delta E\Big[\max_{e'} U(e', z'; \psi, \mu) + \delta E\Big[\max_{e''} U(e'', z''; \psi, \mu) + \cdots \Big]\Big]\Big\},$$

where $\Omega = (\mu, \delta)$ is the set of primitives or structural parameters of the underlying utility function μ and the discount factor δ. The expectation of the value function is taken with respect to both the present and future sales shocks.

4. Estimation

Traditionally, the nested fixed-point algorithm (NFXP) developed by Rust (1987) is used to estimate dynamic models. However, NFXP estimators are computationally burdensome because one has to solve the dynamic program numerically over each guess of the parameter space for every iteration. The two-step estimation first introduced by Hotz and Miller (1993) and extended by Bajari et al. (2007) can serve to reduce the computational burden. In this approach, the model estimation proceeds in two steps. In the first step, we estimate the conditional choice probabilities of choosing a certain action as a flexible nonparametric function of state variables. Then, in the second step, these conditional choice probabilities are used to estimate the structural parameters of the sales agent's utility function.

Until recently, it was believed that the accurate estimation of conditional choice probabilities for an agent is impractical when there is unobserved heterogeneity. Arcidiacono and Miller (2011) propose an expectation-maximization algorithm-based approach to accommodate unobserved heterogeneity in the first step of the two-step estimation procedure. We provide one of the first applications of this approach illustrating the empirical validity of the approach in practical applications. We now discuss the details of the two-step estimation procedure.

4.1. Step 1: Estimating CCPs

In this step, we need to estimate a flexible nonparametric mapping between observable states and the salesperson's actions; this requires a nonparametric model of the monthly effort function, $e_i(z_{it}^E)$, that links effort and states in Equation (1). We model the effort function nonparametrically as a combination of basis functions of the state variables. Thus the nonparametric effort function is

$$e_{it} = \sum_{l=1}^{L} \rho_l(z_{it}^E)\lambda_{il}, \tag{2}$$

where the lth basis function is $\rho_l(z_{it}^E)$. In this application, the lth basis function is the lth-order Chebyshev polynomial.[12]

From Equations (1) and (2), we have the following sales response function to estimate:

$$S_{it} = f(z_{it}^D) + \sum_{l=1}^{L} \rho_l(z_{it}^E)\lambda_{il} + \varepsilon_{it}.$$

For z_{it}^D, which is a subset of z_{it}^E (hereafter referred to as z_{it}), we use two variables: (i) the lagged annual quota for salesperson i and (ii) the ISR index. We use the direct linear effect of these variables to control for cross-sectional variations of territory characteristics and temporal variations in monthly seasonality. The interaction effects of these variables with the other state variables go into the polynomial function in (2).[13]

The lagged annual quota takes into account general territory characteristics that are likely to be generated with limited effort (i.e., market size). The revenues from the indirect sales force capture market seasonality, independent of the nonlinear nature of the compensation plan. We assume that the revenue shocks (ε_{it}) come from an independent and identically distributed normal distribution.[14]

If it were possible to estimate the sales response and effort response functions at the level of each individual, then one could simply obtain the individual-level parameters of the effort and sales policy functions by maximizing the log-likelihood of the sample as

$$\hat{\Theta}_i = \arg\max_{\Theta_i} \sum_{t=1}^{T} \log\left\{ L_i\left(S_{it} - f(z_{it}^D; \alpha_i) - \sum_{l=1}^{L} \rho_l(z_{it}^E)\lambda_{il} \right) \right\}, \tag{3}$$

where the vector $\Theta_i = \{\alpha_i, \lambda_i, \sigma_i\}$ contains the set of parameters of the sales response and effort policy functions and the distribution of sales shocks, where

$$L_i(\varepsilon) = \frac{1}{\sigma_i(2\pi)^{1/2}} e^{-(1/2)(\varepsilon/\sigma_i)^2}. \tag{4}$$

We accommodate unobserved heterogeneity by allowing for discrete segments. Assume that salesperson i belongs to one of K segments, $k \in \{1, \ldots, K\}$ with segment probabilities $q_i = \{q_{i1}, \ldots, q_{iK}\}$. Let the population probability of being in segment k be π_k. Let $\mathcal{L}(S_{it} \mid z_{it}, k; \Theta_k)$ be the likelihood of individual i's sales being S_{it} at time t, conditional on the observables z_{it} and the unobservable segment k, given segment parameters Θ_k. Then the likelihood of observing sales history S_i over the time period $t = 1, \ldots, T$, given the observable history z_i and the unobservable segment k, is given by

$$L_k(S_i \mid z_i; \Theta_k, \pi_k) = \pi_k \left(\prod_{t=1}^{T} \mathcal{L}_{ikt} \right), \tag{5}$$

where $\mathcal{L}_{ikt} = \mathcal{L}(S_{it} \mid z_{it}, k; \Theta_k)$. As noted earlier, we assume the distribution of the revenue shocks to be normally distributed and hence use the normal likelihood for Equation (5) as in Equation (4). The parameter $\Theta_k = \{\alpha_k, \lambda_k, \sigma_k\}$ is the vector of segment-level parameters of the sales response and effort policy functions where each λ_k is the parameter that indexes the effort policy for segment k, and σ_k is the parameter for the distribution of the revenue shocks for segment k.

By summing over all of the unobserved states $k \in \{1, \ldots, K\}$, we obtain the overall likelihood of individual i:

$$L(S_i \mid z_i; \Theta, \pi) = \sum_{k=1}^{K} L_k(S_i \mid z_i; \Theta_k, \pi_k),$$

[12] Chebyshev polynomials are a sequence of orthogonal polynomials, defined by the recurrence relation $T_0(x) = 1$, $T_1(x) = x$, and $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$. By using the orthogonal Chebyshev polynomials to approximate the policy function, rather than approximate the policy function with the standard 1, x, x^2 polynomials, we avoid multicollinearity issues. In our approximation of the effort policy function (reported in Table 6(b)), we use up to third-order Chebyshev polynomials; i.e., $T_0(x) = 1$, $T_1(x) = x$, $T_2(x) = 2x^2 - 1$, and $T_3(x) = 4x^3 - 3x$. For more details on Chebyshev polynomials, we refer the reader to Judd (1998).

[13] Separating sales effects stemming from seasonality or territorial characteristics and effort is a difficult challenge given the unobservability of effort. Our maintained assumption that the main effect of z_{it}^D isolates the territory and seasonal effects from the effort policy function is valid if the firm sets quotas in proportion to territory characteristics and seasonality (z_{it}^D). This remains an important issue to explore in future research.

[14] It is noteworthy that Mirrlees (1999) shows that the first-best can be approximated with normal errors in the sales response function in a one-shot model setting if unbounded punishments are feasible.

and hence the log-likelihood over the N sample of individuals becomes

$$\sum_{i=1}^{N}\log\big(L(S_i \mid z_i; \Theta, \pi)\big)=\sum_{i=1}^{N}\log\bigg(\sum_{k=1}^{K}\pi_k\prod_{t=1}^{T}\mathscr{L}_{ikt}\bigg). \quad (6)$$

Directly maximizing the log-likelihood in (6) is computationally infeasible because the function is not additively separable. So we use the approach of Arcidiacono and Jones (2003) and Arcidiacono and Miller (2011) to iteratively maximize the expected log-likelihood in Equation (7):

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{t=1}^{T} q_{ik}\log\mathscr{L}(S_{it} \mid z_{it}, k; \Theta_k), \quad (7)$$

where q_{ik} is formally defined below as the probability that individual i is of segment type k given parameter values $\Theta=\{\Theta_1, \dots, \Theta_K\}$, where $\Theta_k=\{\alpha_k, \lambda_k, \sigma_k\}$, and segment probabilities $\pi=\{\pi_1, \dots, \pi_k\}$, conditional on all of the observed data of individual i:

$$\Pr(k \mid S_i, z_i; \Theta, \pi) = q_{ik}(S_i, z_i; \Theta, \pi) = \frac{L_k(S_i \mid z_i; \Theta_k, \pi_k)}{L(S_i \mid z_i; \Theta, \pi)}. \quad (8)$$

The iterative process is as follows: We start with an initial guess of the parameters Θ^0 and π^0. A natural candidate for such starting values would be to obtain the parameters from a model without unobserved heterogeneity and slightly perturbing those values.[15] Given the parameters $\{\Theta^m, \pi^m\}$ from the mth iteration, the update of the $(m+1)$th iteration is as follows:

(a) Compute $q_{ik}^{(m+1)}$ using Equation (8) with Θ^m and π^m.

(b) Obtain $\Theta^{(m+1)}$ by maximizing (7) evaluated at $q_{ik}^{(m+1)}$.

(c) Update $\pi^{(m+1)}$ by taking the average over the sample such that

$$\pi_k^{(m+1)}=\frac{1}{N}\sum_{i=1}^{N} q_{ik}^{(m+1)}.$$

We iterate (a)–(c) till convergence.

For the basis functions in the effort policy, we use Chebyshev polynomials of state variables to approximate effort, as described in footnote 12. From the estimation, we obtain the vector of parameters for the basis functions (λ), the vector of parameters for the sales policy (α), and the parameters of the revenue shocks (σ) for each segment k. We also obtain the population segment probabilities (π) for each segment. Thus, this procedure gives us the sales revenue function $\hat{S}(\cdot)$ and effort policy function $\hat{e}(\cdot)$ for each segment. We use these segment-level policies to obtain the structural parameters of each segment. A caveat with the two-step estimation procedure is that the first-stage policy function estimates can be biased (and thus generate bias in the structural parameters) if the state variables in the policy function are correlated with the first-stage errors.

4.2. Step 2: Estimating Structural Parameters

The key idea of the two-step estimation is that in the first stage we observe the agent's optimal actions. Using these observed optimal actions, we are able to construct estimates of the value function, which enables us to estimate the primitives of the model that rationalize these optimal actions.

Let the value function of a representative agent at state z that follows an action profile e—conditional on the compensation plan ψ, the sales profile S, and the primitives of the utility function and discount parameters $\Omega=(\mu, \delta)$—be represented as

$$V_t(z; e; \psi, S, \Omega) = E\bigg\{\sum_{t=0}^{T} D(t)U(e(z_t), z_t, \varepsilon_t; \mu)\,\bigg|\, z_0=z; \psi, S, \Omega\bigg\}, \quad (9)$$

where $D(t)=\delta^t$ is the discount factor, and the expectation operator would be over the present and future sales shock ε_t.

Using the estimated sales and effort policy function and the distribution of the sales shocks in the first stage, we are able to forward-simulate the actions of sales agents to obtain the estimate of the value function. The detailed simulation procedure is as follows:

(a) From initial state of z_t, calculate the optimal actions as $e(z_t)$.

(b) Draw sales shock ε_t from $f(\varepsilon)$.

(c) Update state z_{t+1} using the realized sales $s(e(z_t))+\varepsilon_t$.

(d) Repeat (a)–(c) until $t=T$.

By averaging the sum of the discounted stream of utility flow over multiple simulated paths, we can get the estimate of the value function, $\tilde{V}(z; e(z); \psi, S, \Omega)$.[16]

Let $e^s(z)$ be any deviation policy from a set of feasible policies that is not identical to the optimal policy, and, by using the same simulation method proposed above, let the corresponding estimate of the value function be called the suboptimal value function, $\tilde{V}(z; e^s(z); \psi, S, \Omega)$. Since $e(z)$ by definition is the effort policy and thus at an optimum, then any deviations from this policy rule would generate

[15] We started the initial values from one-tenth of the standard error from the parameter values obtained from a single-segment model. The initial values of the segment probabilities were set equally across segments.

[16] For each segment, we drew 400 simulation draws over each period and computed the value functions.

value functions of lesser or equal value to that of the optimal level.

Let us define the difference in the two value functions as

$$Q(v;\psi,S,\Omega) = V(z;e(z);\psi,S,\Omega) - V(z;e^s(z);\psi,S,\Omega),$$

where $v \in \mathcal{V}$ denotes a particular $\{z, e^s(z)\}$ combination.[17] Then if $e(z)$ is the optimal policy, the function $Q(v;\psi,S,\Omega)$ would always have a value greater than or equal to zero. Thus our estimate of the underlying structural parameters Ω would satisfy

$$\hat{\Omega} = \arg\min \int \left(\min\{Q(v;\psi,S,\Omega),0\}\right)^2 dH(v),$$

where $H(v)$ is the distribution over the set $\mathcal{V}$ of inequalities. Our empirical counterpart to $Q(v;\psi,S,\Omega)$ would be

$$\tilde{Q}(v;\psi,\hat{S},\Omega) = \tilde{V}(z;\hat{e}(z);\psi,\hat{S},\Omega) - \tilde{V}(z;\hat{e}^s(z);\psi,\hat{S},\Omega).$$

As a result, our estimates of the structural parameters are obtained from minimizing the objective function in Equation (10):[18]

$$\frac{1}{N_I}\sum_{j=1}^{N_I}\left(\min\{\tilde{Q}(v_j;\psi,\hat{S},\Omega),0\}\right)^2. \tag{10}$$

The above procedure is performed for each segment with the segment-specific effort policies obtained in Step 1. This allows us to estimate the structural parameters for each segment.[19] In practice, since the objective function in Equation (10) is relatively flat with respect to changes in the discount factor, it is difficult to pin down the discount parameter with traditional gradient-based optimization. In our specification, we first estimate the structural parameters by minimizing the objective function in Equation (10) over a grid of discount factors. Second, we simulate data with the estimated structural parameters associated with each discount factor to compute the mean absolute percentage error (MAPE) per period with respect to the observed data. We choose the final estimates as the discount factor and the associated structural parameters with the lowest MAPE.[20]

4.3. Identification

There are a couple of major identification challenges. First, we do not observe effort. Hence the link between effort and sales cannot be identified nonparametrically. Second, in dynamic structural models, it is usually impossible to identify discount factors separately from the utility function. Below, we discuss how we address these issues.

Realized sales are a function of demand shifters, effort, and additive sales shocks. Conditional on observed demand shifters and given multiple observations of sales at different states, we can separately identify nonparametrically the density of sales shocks and a deterministic function of effort. We assume a deterministic (but flexible) relationship between effort and observable states (%QQ and %AQ and demand shifters) for each segment. Finally, because we do not observe effort, we need a strictly monotonic parametric relationship between sales and effort. As we estimate a flexible relationship between observable states and effort, we model the relationship between sales and effort to be linear.

We rely on the variation in states across agents and within agents for identification. The average coefficient of variation of states %AQ within agents (i.e., $\text{CV} = 1/N\sum_{i=1}^{N}\sigma_i^{\%AQ}/\mu_i^{\%AQ}$, where μ_i^{AQ} and σ_i^{AQ} are the mean and standard deviation, respectively, of the distance to annual quotas over three years for agent i) over the three years is 0.55. This allows us to separate effort from the individual fixed effects. The average coefficients of variation across agents and months on %AQ and %QQ are 0.63 and 0.91, respectively.[21] The variation in states across agents helps separate agent effort from the monthly fixed effects.

The discount factor is not identified separately from the utility function in standard dynamic structural models because, typically, there are no variables that do not affect contemporaneous utility, only future

[17] As indicated in Bajari et al. (2007), there are multiple ways to draw these suboptimal policy rules. Although the method of selecting a particular perturbation will have implications for efficiency, the only requirement necessary for consistency is that the distribution of these perturbations has sufficient support to yield identification. We chose to draw a deviation policy from a normal distribution with mean zero and a quarter of the variance from the revenue shock distribution; i.e., $e^s(z) = e(z) + \eta$.

[18] We drew 200 deviation strategies to construct the objective function, and hence $N_I = 200$.

[19] In addition, we used a second set of moment inequalities to reflect the participation constraint that employees continued to work at a firm because they at least obtained a reservation value (normalized to zero); i.e., $\min\{V(z;e(z);\psi,S,\Omega),0\}$. It turns out these inequalities are nonbinding and do not affect our estimates.

[20] The statistical properties of the grid-search estimator based on MAPE for discount factors are not known. Hence, we computed the standard errors using subsample bootstrap methods, where we subsampled the data and estimated the model multiple times to construct a bootstrap distribution (Efron 1979). In her analysis of entry and exit in the washing machine industry in the United States, Shen (2013) also uses grid search to estimate the discount factor of firms; she uses parametric bootstrapping to estimate the standard error.

[21] There is substantial variation in the across-agent CV of %AQ and %QQ across months.

utility (Rust 1994, Magnac and Thesmar 2002). In the absence of such an exclusion restriction, this implies that if an agent exerts low effort in a period, it is not possible to distinguish whether this is due to high disutility for effort or because she discounts future utilities very heavily.[22]

Two aspects of our setting allow us to identify utility functions separately from the discount factor. First, we have a finite-horizon setting, where at the end of the year, the quotas are reset and all agents start with a fresh quota for the following year. This means that every December, the agent faces a static optimization problem, conditional on the sales agent's state (%AQ). Utility parameters are well identified for a static model, and hence the agent's choice in the last period should allow us to nonparametrically identify the agent's utility function. Given this, a variation in sales (that is monotonically linked to effort) in the last period and variations in wealth should help identify the effort disutility and risk-aversion coefficient within the utility function.

Second, the bonus setting generates exclusion restrictions between the current and future utility; i.e., we have instruments in nonbonus periods that do not affect current utility, only future utility. As we demonstrated with reduced-form evidence earlier, that an agent's performance in November is related to his proximity to the annual bonus given in December indicates forward-looking behavior. We also collected other evidence of how agents respond to quarterly or annual quotas even though they do not affect current payoffs. This allows us to estimate the discount factor.

Beyond these conceptual arguments, we report in the appendix the results of a simulation study that demonstrate that the structural parameters of the utility function and discount factor can be identified using our empirical strategy.

5. Results

We first report the first-stage estimates of the demand shifters and effort policy function for the sales response model; then we report estimates of structural parameters of sales agents' utility functions from the second-stage estimation. We then perform several counterfactual simulations to address the substantive questions we seek to answer.

[22] Another reason why an agent might exert low effort is that she may have incorrect expectations about the transition density of future states; i.e., she might be very pessimistic about future good states. Similar to other dynamic structural modeling papers, we assume rational expectations for the transition densities of states. In this case, this translates into a rational expectations assumption on sales shocks.

5.1. First-Stage Estimates

The parameter estimates for the demand shifters in the sales response function is reported in Table 6(a). We find that lagged annual quota and the interaction term between it and indirect sales revenue are statistically significant. Thus larger markets tend to have a bigger sales multiplier independent of effort in high-demand periods.

We estimate segment-level effort policy functions by estimating the nonparametric relationship between sales and state variables through Chebyshev polynomials of the state variables. We estimated up to fourth-order Chebyshev polynomials with an alternative number of segments and chose the best-fitting model based on the Bayesian information criterion. The best fit was for the model with up to third-order Chebyshev polynomials, allowing for three segments. The estimates of the best-fitting polynomial function and the standard deviations of the revenue shocks for each segment are reported in Tables 6(b) and 6(c). As the coefficients associated with the Chebyshev polynomials have no intuitive meaning, for intuition, we show graphs of the effort policy function for the three segments as a function of the percentage annual quota (%AQ) and percentage quarterly quota (%QQ) for select months in Figure 6(a). Both %AQ and %QQ are normalized across sales agents, such that 1 implies meeting quota and 0.9 indicates 10% below quota and 1.1 indicates 10% above quota.

Table 7 shows the share of the three segments and their descriptive characteristics. Segment 2 is the largest with a share of 47%; Segments 1 and 3 have shares of 32% and 21%, respectively. The average tenure with the firm is not very different across segments, at approximately 12 years. Segment 3 has the highest annual quotas, followed by Segments 2 and 1. It is worth noting that Segments 2 and 3, with larger quotas, achieve their quota targets more often than Segment 1, which has trouble meeting its quota.

Figure 6(a) shows that Segment 3 exerts the most effort and is the most productive segment, and Segment 1 exerts the least effort and is the least productive segment. This is consistent with the allocated quotas and percentage of time quotas are achieved in Table 7. We also see a positive relationship between exerted effort and %AQ for all months shown. As for

Table 6(a) Parameter Estimates—Sales Response

Lagged annual quota	-0.022^{***} (0.007)
Indirect sales	-0.955 (6.307)
Indirect sales $\times$ *Lagged annual quota*	0.023^{***} (0.003)

Note. Standard deviations are shown in parentheses.
$^{***}p < 0.01$.

Table 6(b) Parameter Estimates—Effort Policy Function

Variable	Segment 1	Segment 2	Segment 3	Variable	Segment 1	Segment 2	Segment 3
ρ_0	-138.25^{***}	-94.17^{**}	-214.82^{*}	$\rho_1(z_2)\rho_1(z_3)$	2.74	196.33^{***}	$1{,}011.25^{***}$
	(37.81)	(45.72)	(116.77)		(37.73)	(49.87)	(138.31)
$\rho_1(z_1)$	59.15	163.96	308.24	$\rho_1(z_2)\ \rho_1(z_2) \times \rho_1(z_3)$	-56.61	-168.63^{***}	-101.05
	(78.25)	(101.85)	(248.03)		(53.89)	(64.36)	(154.00)
$\rho_2(z_1)$	-32.43	-42.48	-201.92^{***}	$\rho_1(z_4)$	-7.44^{***}	-3.85^{**}	-2.53
	(22.06)	(28.76)	(77.66)		(1.16)	(1.66)	(3.74)
$\rho_3(z_1)$	6.53	11.25^{*}	27.30^{*}	$\rho_1(z_1)\rho_1(z_4)$	14.48	-2.68	-13.97
	(4.68)	(5.99)	(15.22)		(15.94)	(21.58)	(58.56)
$\rho_1(z_2)$	259.22^{***}	35.67	-685.74^{***}	$\rho_1(z_2)\rho_1(z_4)$	15.54^{***}	26.00^{***}	108.23^{***}
	(63.82)	(79.01)	(195.49)		(4.41)	(5.27)	(13.48)
$\rho_2(z_2)$	-146.79^{***}	-85.54^{**}	-61.03	$\rho_1(z_2)\ \rho_1(z_2) \times \rho_1(z_4)$	-15.44	-17.99	54.90
	(30.45)	(34.51)	(78.18)		(13.56)	(14.82)	(37.56)
$\rho_3(z_2)$	28.72^{***}	12.55^{**}	8.90	$\rho_1(z_1)\ \rho_1(z_3) \times \rho_1(z_4)$	-4.60	8.41	10.96
	(5.89)	(6.05)	(13.37)		(5.54)	(7.54)	(20.43)
$\rho_1(z_1)\rho_1(z_2)$	74.25	151.11	-349.73	$\rho_1(z_2)\ \rho_1(z_3) \times \rho_1(z_4)$	-0.25	-20.80^{***}	-102.95^{***}
	(167.36)	(199.47)	(490.91)		(3.87)	(5.17)	(14.63)
$\rho_1(z_1)\rho_2(z_2)$	1.47	82.06^{***}	-58.59	$\rho_1(\tau)$	-0.16	8.17^{***}	15.58^{**}
	(35.90)	(31.81)	(68.07)		(1.38)	(2.12)	(6.81)
$\rho_2(z_1)\rho_1(z_2)$	16.51	-24.49	49.78	$\rho_2(\tau)$	0.03	-0.24^{***}	-0.41
	(20.83)	(24.10)	(53.17)		(0.05)	(0.07)	(0.26)
$\rho_1(z_1)\rho_1(ISR)$	-37.24^{***}	-65.19^{***}	-57.22	$\rho_3(\tau)$	0.00	0.00^{***}	0.00
	(12.68)	(15.98)	(40.65)		(0.00)	(0.00)	(0.00)
$\rho_1(z_2)\rho_1(ISR)$	-13.25^{***}	-0.61	-38.38^{***}	$\rho_1(z_1)\rho_1(AQ_{y-1})$	-0.04^{**}	-0.03^{**}	-0.07^{***}
	(3.35)	(3.91)	(10.27)		(0.02)	(0.02)	(0.03)
$\rho_1(z_2)\ \rho_1(z_2) \times \rho_1(ISR)$	37.42^{*}	47.50^{**}	2.84	$\rho_1(z_2)\rho_1(AQ_{y-1})$	-0.01	-0.06^{***}	-0.02^{*}
	(21.48)	(23.67)	(53.51)		(0.01)	(0.01)	(0.01)
$\rho_1(z_3)$	-0.16	-10.00^{**}	-43.33^{***}	$\rho_1(z_2)\ \rho_1(z_2) \times \rho_1(AQ_{y-1})$	0.04	0.06^{*}	0.03
	(3.39)	(5.10)	(11.92)		(0.04)	(0.03)	(0.04)
$\rho_1(z_1)\rho_1(z_3)$	54.44^{**}	81.66^{**}	113.02				
	(26.97)	(37.01)	(97.94)				

Note. Standard deviations are shown in parentheses.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

%*QQ*, we see an increasing but concave relationship in March, implying that once a salesperson is way above the quarterly quota, she starts to gradually slow down. Given that the average states in March for each segment were 0.55, 0.58, and 0.62, respectively, not a lot of salespeople are in a position where they can slow down. Effort in December does not fall off even if the salesperson has already reached or exceeded her quota ($\%AQ > 1$), likely because of the overachievement commissions in preventing salespeople from lowering effort after achieving quotas. Our results are consistent with Steenburgh (2008), who finds that salespeople "give up" when far away from achieving their quota, such as for all segments in our case, but they do not slow down much once their quota is reached.

Figure 6(b) shows the effect of tenure on effort for all segments. Salespeople in Segments 2 and 3 initially increase effort with experience, but this tapers off with time. This is probably because in the early years of their careers, they want to work hard not only for monetary payments from increased wages but also for other intangible incentives, such as promotions or transfers to better job titles. However, after a certain number of years, these intangibles do not matter as much, and the effort levels tend to taper off. It is noteworthy that Segment 1, the lowest-productivity segment, does not gain in productivity from experience.

Table 6(c) Revenue Shock Distribution

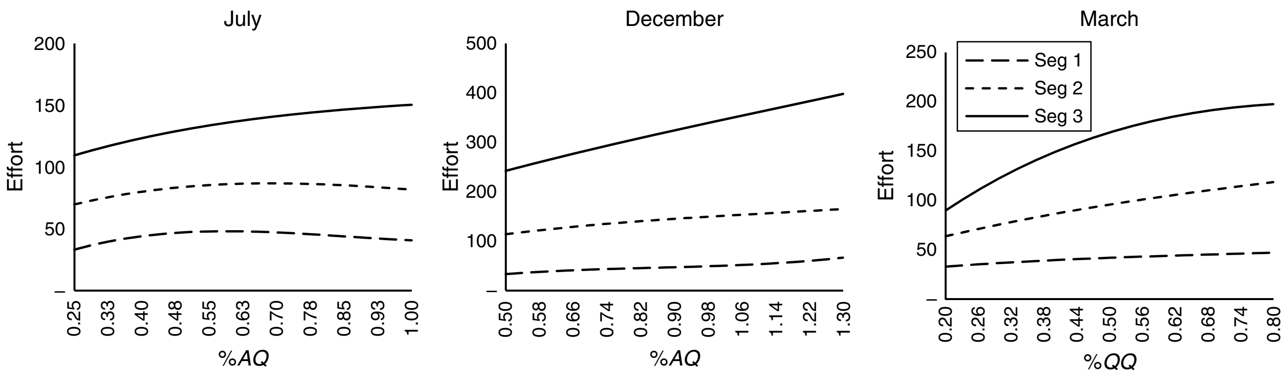
	Segment 1	Segment 2	Segment 3
Sigma	80.59	141.84	271.73

Table 7 Descriptive Characteristics of Segments

	Segment 1	Segment 2	Segment 3
Share	0.32	0.47	0.21
Tenure[a]	11.21	12.34	11.60
Achieve quarterly quota: Q1	0.46	0.54	0.57
Achieve quarterly quota: Q2	0.38	0.55	0.62
Achieve quarterly quota: Q3	0.30	0.49	0.53
Achieve annual quota	0.30	0.57	0.65
Average annual quota[b]	1,199.08	1,623.11	2,349.10
Average December revenue[b]	127.30	272.53	564.64

[a]Tenure is measured in years.
[b]Average quotas and revenues are indicated in U.S. dollars ('000).

Figure 6(a) **Effort Policy by Segment as a Function of Percentage of Quota**

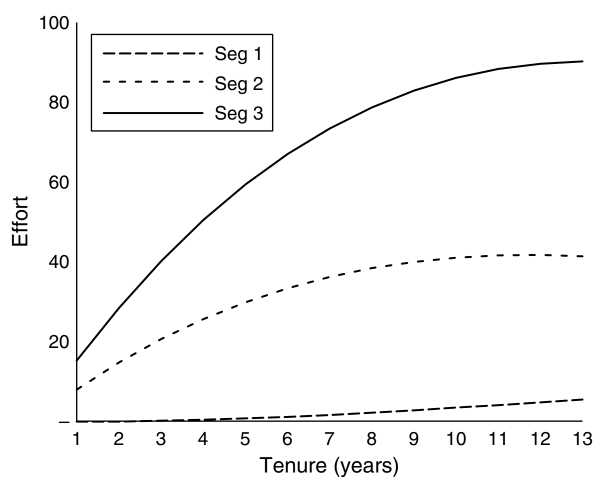


5.2. Second-Stage Structural Parameter Estimates

The first column of Table 8 shows the estimates of the dynamic model with exponential discounting. We find the discount factor δ to be 0.9 and highly significant. Frederick et al. (2002) have a comprehensive summary of the estimated discount factors from previous studies. The summary shows that the estimated discount factors vary extensively, ranging from as low as a mere 0.02 to no discounting at all, with a discount factor of 1. For purely monetary values, the estimated discount factor seems rather low. But as Frederick et al. point out, for behavioral aspects such as pain—and in our case, effort—the discount factors tend to be low. Hence, our estimate appears to be reasonable.

The parameters for disutility of effort are negative and significant for all three segments. Their relative magnitudes are consistent with the effort policy functions estimated in the first stage. Segment 3, which produces the greatest sales on average, has the lowest disutility for effort. Segment 1, which has the lowest sales, has the greatest disutility. The risk-aversion coefficients for all segments are insignificant, showing no direct evidence of risk aversion by the sales agents. This may be because in the range of incomes earned by the sales force, risk aversion is not a serious concern. The estimated model fits the observed sales revenue data reasonably well with a MAPE of 9.4%.

Figure 6(b) **The Effect of Tenure on Effort**



5.3. Assessing the Value of a Dynamic Structural Model

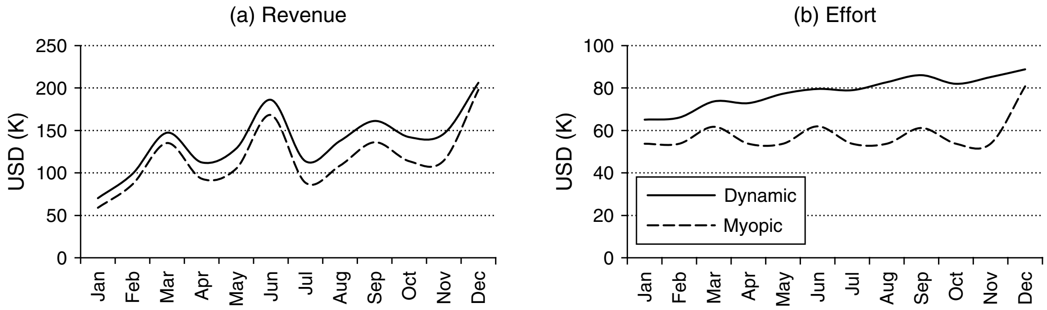
How important is it to model the dynamics of salesperson behavior? In a static model, any effort would be attributed to the current payoff, not accounting for the large future bonuses. This will downward bias the salesperson's disutility parameters and overstate the effects of compensation on productivity. The second column of Table 8 reports the estimates of the myopic model, where the discount factor is set to zero. As expected, the disutility parameters are smaller in magnitude relative to the forward-looking model for all segments. For Segment 3, the downward bias is as much as 28%. The myopic model also has a

Table 8 **Structural Parameters**

	Dynamic	Myopic
Discount factor	0.90***	
	(0.04)	
Segment 1		
Disutility	−0.375**	−0.284**
	(0.189)	(0.114)
Risk aversion	−0.0001	−0.0001
	(0.0020)	(0.0003)
Segment 2		
Disutility	−0.145***	−0.111***
	(0.038)	(0.032)
Risk aversion	−0.0001	−0.0001
	(0.0005)	(0.0005)
Segment 3		
Disutility	−0.077***	−0.056***
	(0.016)	(0.010)
Risk aversion	0.0000	0.0000
	(0.0002)	(0.0004)
MAPE	0.094	0.167

Note. Standard errors are shown in parentheses.
** $p < 0.05$; *** $p < 0.01$.

Figure 7 Simulated Revenue and Effort—Static vs. Dynamic



poorer fit: a MAPE of 16.7% relative to MAPE of 9.4% for the dynamic model.

We next compare the revenue and effort predictions between the dynamic and myopic models. To isolate the effects of forward-looking behavior, we simulate based on the disutility estimates from the dynamic model but set the discount parameter to zero for the myopic model. Figure 7 compares the predicted revenues and effort of the myopic and dynamic models. The myopic agent has systematically lower revenues because she does not take into account the effect of future bonuses and overachievement commission in current effort. In contrast, the forward-looking agent anticipates that in an uncertain environment, there is a chance of bad shocks later that may prevent her from meeting the quota, so she prepares for such a rainy day by working harder early on so that she is within striking distance of her quota even if a bad sales shock occurs.

The effort graph in Figure 7 enables us to isolate out the sales revenue cyclicality and focus on the differences in effort across dynamic and myopic agents. The myopic salesperson concentrates much more effort in the bonus period, but the forward-looking salesperson smooths effort over time, given the uncertainty in future demand shocks. The effort peaks in the bonus periods are not as pronounced for the dynamic agent. The observed effort smoothing is similar to consumption smoothing by forward-looking consumers facing uncertain incomes in the development economics literature.

5.4. Counterfactual Simulations

We now perform a series of counterfactual simulations that address the two sets of substantive questions we wish to answer. First, we address the issue of how valuable different components of the compensation plan are. The overall change in revenues under the alternative conditions is reported in Table 9 and the effect by segment in Table 10. Second, we compare the role of bonus frequency—how quarterly and annual bonuses affect performance. In running the counterfactual simulation studies, given a regime change, we randomly chose 25 representative sales agents from each segment (a total of 75 salespeople) and simulated 1,000 simulation paths for each of them to obtain the productivity of each segment. Then, we weighted each segment based on the estimated segment sizes reported in Table 7.

5.4.1. Value of Nonlinear Incentive Components. We compare changes in revenues and profits when the firm moves from the current compensation plan to a pure commission-only plan. We consider two cases: (1) where the commission rate is the same as the current commission rate and (2) where a higher commission rate is such that total compensation is exactly equal to the current compensation. We find that the revenues are about 17.9% greater with the current compensation plan compared with a pure commission plan. Not surprisingly, given the lower incentives in the absence of quotas, bonuses, and overachievement commissions, all segments suffer from substantially poorer performance, as shown in Table 10. Even after adjusting commission rates to be higher to make total compensation identical to the current plan, revenues are 2.4% higher with the current compensation plan, suggesting that the nonlinear plan shifts salespeople into high-power areas of incentives. To understand the role of different components, we next investigate counterfactuals based on individual components one at a time.

Table 9 Impact of Alternative Bonus Plans on Sales Revenues

Counterfactual	Change in revenues (%)
Only pure commissions	−17.9
Only pure commissions (adjusted to equal payout with bonus)	−2.4
No bonus (only commissions + overachievement commission)	−8.0
No bonus (commissions adjusted to equal payout with bonus)	−2.6
No overachievement commissions	−10.7
Cumulative annual quota replaced with quarterly quota	−1.7
Annual bonus split into quarterly and annual bonus	−0.5
Remove quarterly bonus	−3.9

Table 10 Impact of Alternative Bonus Plans on Sales Revenues by Segment

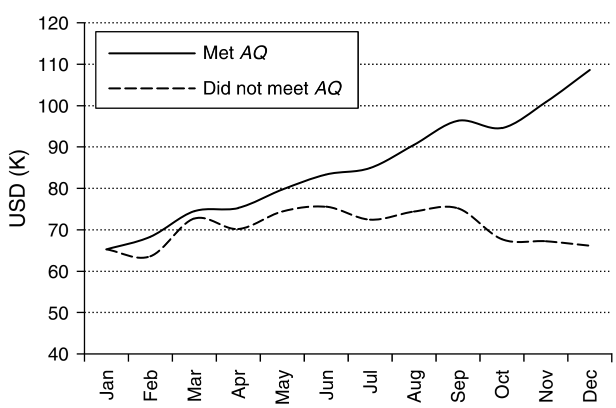
Percentage of decrease from different components	Segment 1 (%)	Segment 2 (%)	Segment 3 (%)
Pure commission	10.5	19.4	19.3
Without overachievement	2.4	10.8	14.5
Without quarterly bonus	6.0	4.4	2.2

5.4.2. Value of Overachievement Compensation. When overachievement commissions are eliminated, not surprisingly, overall revenues drop by 10.7%. Even accounting for the additional commission costs, profits are lower by about 2% (assuming a gross margin of 33%).

To gain insight into how overachievement commissions impact sales, Figure 8(a) compares the effort level of sales agents who eventually meet or do not meet the annual quota. For those who met the annual quota, their effort level does not decline even when close to the quota because of the overachievement commission. In contrast, those who did not meet the annual quota decrease effort toward the end of the year because they are unlikely to meet the quota, and therefore the overachievement commission has no impact on their earnings. Thus the overachievement commission provides incentives for the most productive salespeople even if they have already met their quota (or are likely to meet it). Table 10 indicates, not surprisingly, that overachievement commissions have the greatest impact on Segment 3, the most productive segment. Revenues drop by about 15% for this segment, whereas the effect on the least productive segment is substantially smaller, at 2%. Overall, we conclude that overachievement commission rates motivate high-performing salespeople to continue exerting high levels of effort when they are close to and exceed the quota, but these rates have a limited impact on low-performing salespeople.

5.4.3. Value of a Cumulative Annual Quota. Next we investigate the effect of cumulative annual quota by replacing the cumulative annual quota with just a fourth-quarter quota. That is, we set the fourth-quarter quota as the annual quota minus the sum of three previous quarters' quotas ($QQ_4 = AQ - QQ_1 - QQ_2 - QQ_3$) and split the total bonus payments (annual bonus + quarterly bonuses) equally across all four quarters.[23] Furthermore, to isolate the effect of bonuses, we also remove the overachievement commission for reaching the annual quota in both scenarios. Overall, revenues drop by 12.0% when the cumulative annual quota and overachievement commissions are removed. This decrease is greater than the 10.7% we obtained when we just dropped the overachievement commission. Thus the cumulative annual quota induces sales agents to exert greater effort and raise revenues by 1.3%.

Figure 8(a) Effort: Overachievement Commission

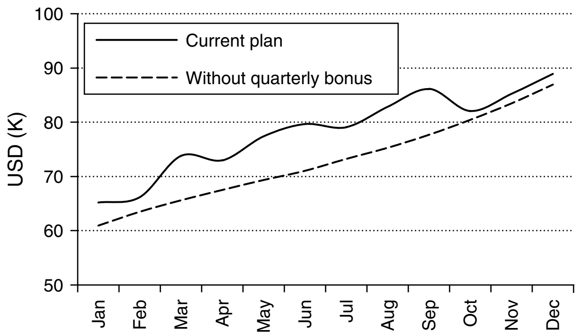


How does the cumulative annual quota work? On the negative side, the cumulative annual quota reduces the incentive of salespeople who have not had success in the first three quarters to continue to put in effort in the fourth quarter. But on the positive side, the cumulative annual quota allows salespeople who reach their quarterly quota in earlier quarters to continue to extend themselves in order to get them closer to the annual target—making this additional effort early on helps sales agents buffer against negative shocks in future quarters and allows them to be within striking distance of the annual quota. On balance, the positive effect of the latter outweighs the negative effect of the former.

We also consider the case where we split the current annual bonus into a fourth-quarter bonus (the same amount as other quarterly bonuses) and an annual bonus so that salespeople do not give up in the last quarter when they are far away from quota. Specifically, we split the $4,000 annual bonus into a $1,500 fourth-quarter bonus and a $2,500 annual bonus. Although this did increase the effort in the last quarter, it reduced revenues overall because salespeople did not put in as much effort earlier in the year to be within striking distance of the annual quota because it is not as large. Total revenues dropped by 0.5%.

[23] Specifically, the quarterly bonus is set to be (annual bonus + 3 × quarterly bonus)/4; i.e., $(4{,}000 + 1{,}500 \times 3)/4 = \$2{,}125$.

Figure 8(b) **Effort: Quarterly Quotas**



Overall, these results show that the cumulative annual quota has a significant impact on salespeople by inducing agents to work harder even when they reach their quarterly targets, again motivating the high-performing salespeople who reach quarterly quotas to continue to remain productive and bring additional sales.

5.4.4. Quota-Bonus Frequency. We next investigate the value of quarterly bonuses relative to annual bonuses. Figure 8(b) shows the comparison of effort between the current plan and when quarterly bonuses are eliminated and only the annual bonus is left. Effort drops consistently across the year when there are no quarterly quotas. Overall revenues fall by 4%. Even in December, when the annual bonus is on the table, revenue falls by 1% and effort falls by 2%. Thus annual bonus and overachievement commissions have less of an impact on year-end performance without quarterly bonuses. Why?

The quarterly bonus induces sales agents to work harder in a given quarter. But it also helps them achieve the annual quota by helping them stay on track of their annual goal. Without a quarterly bonus, sales agents do not have much incentive to work hard early on. This lack of incentive leads them to be farther away from the annual quota by December. Annual bonuses and overachievement commission have little impact on effort as sales agents are more likely to give up meeting the annual quota.

The impact of quarterly bonuses also differs across the three segments of consumers. Table 10 indicates that quarterly bonuses have a relatively minimal impact on Segment 3, the most productive segment, but they almost triple the impact of Segment 3 on Segment 1. In effect, quarterly bonuses are needed as pacers for the less productive salespeople than for the most productive salespeople.

To the best of our knowledge, there has been no analysis to date on what the appropriate frequency of quota and bonuses is. There has been some descriptive work in the education literature on how frequent testing affects academic performance (for an extensive survey, see Bangert-Drowns et al. 1991) and some experimental work in behavioral psychology (Heath et al. 1999). The basic idea is that achieving short-term goals makes achieving long-term goals more feasible. Our analysis shows that the short-term goals are more valuable to the least productive segment; i.e., in education terms, it implies that weaker students gain more by periodic testing relative to stronger students who would study regardless of exams.

5.4.5. Discussion. Overall, our results provide empirically grounded substantive insight on the role of various elements of nonlinear compensation to improve productivity. The numerical example in §2.2 shows that quotas and bonuses serve as important goals for the average performers, inducing them to increase their effort. Our empirical analysis shows support for the role of quotas and bonuses as motivational goals and stretch incentives, respectively. However, it is important to recognize that other elements of the compensation plan are critical in improving the performance of both high performers and low performance. First, in this setting, managers recognize that quota updating based on the salesperson's own performance in the previous year can induce ratcheting effects; such effects can reduce productivity among the highest-performing salespeople, who do not book sales much higher than their quota so as to avoid higher quotas the following year. By updating quotas based not on an individual's performance but on the performance of a larger group of salespeople, managers can minimize ratcheting effects. In fact, we found empirically little statistical evidence of effort shading as a result of ratcheting effects. Second, the high performers who are most likely to reach their quotas will reduce effort upon reaching them if there are no additional incentives offered. The firm used overachievement commissions and cumulative annual quotas to induce the high-performing sales agents to continue to put in effort and excel in their sales performance. These incentives, along with the avoidance of ratcheting effects, enabled the firm to keep its best sales agents motivated to continue performing at a high level. Finally, quarterly bonuses served as pacers to help keep the least productive segment motivated so that it can be within striking distance of its long-term goals. By providing empirically grounded insight into the differential motivational roles of different incentive components of a nonlinear compensation plan for high-, average-, and low-performing sales agents, this paper contributes substantively to literature on the role of nonlinearities in quota-bonus plans and explains its widespread popularity in practice (Joseph and Kalwani 1998).

6. Conclusion

Personal selling is a primary marketing mix tool for most B2B firms to generate sales, yet there is little research on how the compensation plan motivates the sales force and affects performance. This paper develops and estimates a dynamic structural model of sales force response to a compensation plan with various components: salary, commissions, lump-sum bonuses for achieving quotas, and different commission rates beyond achieving quotas. Our analysis helps us assess the impact of (1) different components of compensation and (2) the differential importance of periodic bonuses on the performance of different types of salespeople.

We find that the quota-bonus scheme used by this firm increases the sales force's performance by serving as an intermediary goal and pushing employees to meet targets. Features such as overachievement compensation reduce the problems associated with sales agents slacking off when they get close to achieving their quota. Furthermore, quarterly bonuses serve as a continuous evaluation scheme to keep sales agents within striking distance of their annual quotas. In the absence of quarterly bonuses, failure in the early periods to meet targets cause agents to fall behind more often than in the presence of quarterly bonuses. Thus, the quarterly bonus serves as a valuable subgoal that helps the sales force stay on track in achieving its overall goal; such incentives are especially valuable to low performers. In contrast, overachievement commissions increase performance among the highest performers.

In this empirical application, we introduce two important methodological innovations to the marketing literature. First, to the best of our knowledge, we provide the first empirical implementation of the Arcidiacono and Miller (2011) approach to accommodate unobserved heterogeneity within a two-step estimation framework. Second, we demonstrate that discount factors can be estimated in naturally occurring field data using appropriate exclusion restrictions—and we overturn the conventional wisdom that estimation of discount factors requires augmentation with survey data.

We now discuss the limitations of our paper, which provide promising avenues for future research. First, effort tends to be multidimensional, and one possibility is that quotas and bonuses force people to focus on the effort that leads to final sales in bonus periods, whereas agents may focus on the earlier stages of the selling process in nonbonus periods. Such a multidimensional effort cannot be identified merely from sales data. We hope data from customer relationship management databases that track customer stages through the selling process can help shed light on this issue. We believe this is an exciting area for future research.

Second, compensation contracts can serve to select the right type of salespeople. We do not address selection issues. One possibility is to use a longer panel of salespeople's performance that includes attrition information. If there were variation in contracts that affected employee retention, that could also help address this problem. More work on scenarios with richer contracts needs to be done. For example, one could study peer effects on sales performance and selection effects when firms shift from individual- to team-based compensation (Chan et al. 2014).

In summary, this paper provides a rigorous framework to empirically understand how the sales force responds to a very rich compensation structure involving many components of compensation: salaries, commissions, quotas, and bonuses at quarterly and annual frequencies. Our analysis helps obtain a number of useful substantive insights. Nevertheless, the issues raised above provide an interesting agenda for future work.

Acknowledgments
This paper is the first essay of the first author's doctoral dissertation at Yale University. The authors thank the editor, associate editor, two anonymous reviewers, Lanier Benkard, Adam Copeland, David Godes, Ahmed Khwaja, Subrata Sen, as well as the seminar participants at Carnegie Mellon University, Cornell University, Emory University, the Georgia Institute of Technology, Harvard University, the Hong Kong University of Science and Technology, the Indian Institute of Management, the Indian School of Business, New York University, Northwestern University, Pennsylvania State University, Stanford University, the University of Arizona, the University of British Columbia, the University of California at Berkeley, the University of Chicago, the University of Houston, the University of Maryland, the University of North Carolina, the University of Rochester, Washington University, Western University, and Yale University, in addition to the conference participants at the 2009 UT Dallas FORMS Conference, the 2010 Choice Conference in Key Largo, the 2010 Sales Conference in Kansas, the 2012 Harvard Thought Leadership on the Sales Profession Conference, and the 2012 INFORMS Marketing Science Conference for their comments and suggestions. The authors also thank ISBM for providing financial support.

Appendix. Simulation Study to Identify Discount Factors

Let realized sales be a function of effort, a function of demand shifter z_t, and a normal random sales shock ε_{it} such that

$$y_{it} = h(z_t) + e_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \sim \mathrm{N}(0, \sigma^2).$$

To illustrate identification, we consider a simulation scenario with seasonality and quarterly quotas. The demand shifter z_t represents the seasonality index comparable to the ISR index in our empirical setting. For the purpose of the simulation, we assume the function $h(\cdot)$ is linear in the seasonality index; i.e., $h(z_t) = \alpha z_t$. Let the salesperson's utility be given as $u_{it} = -de_{it}^2 + W_{it}$, where $W_{it} = BI_{\{t=3,6,9,12\}} I_{\{s_{it}+y_{it}>Q\}}$ is the wealth earned from bonus

Table A.1 Simulation Results

No. of individuals	True values		
	$d = -0.1$	$\delta = 0.9$	$\sigma = 10$
50	−0.134	0.892	10.282
	(0.166)	(0.287)	(0.320)
100	−0.106	0.899	10.266
	(0.025)	(0.006)	(0.221)
200	−0.108	0.899	10.043
	(0.018)	(0.003)	(0.150)
300	−0.105	0.899	9.944
	(0.010)	(0.002)	(0.121)

Note. Standard errors are in parentheses.

payment B for achieving quarterly quota Q. The cumulative sales state within quarter evolves as follows:

$$s_{it} = \begin{cases} 0 & \text{if } t \text{ is the start of quota period,} \\ s_{i(t-1)} + y_{it} & \text{otherwise.} \end{cases}$$

We report the results of the simulation with seasonality and quarterly quotas using three years of data. Let the vector of the seasonality index for each month throughout the year be given by $(2, 3, 5, 1, 2, 3, 3, 4, 7, 2, 3, 5)$. We set the bonus at $B = 60$ and quarterly quota at $Q = 30$. Consistent with the model in the paper, we estimate the first-stage estimates using our specification in Equation (1).

We varied the simulated number of individuals from 50 to 300. The true values and the estimates and standard errors for each simulation are reported in Table A.1. Even with 100 individuals (relative to our sample size of 350), we are able to recover the disutility parameter and the discount factor with reasonable precision, lending confidence to the identification arguments in §4.3.

References

Albers S, Mantrala MK, Sridhar S (2010) Personal selling elasticities: A meta-analysis. *J. Marketing Res.* 47(5):840–853.

Arcidiacono P, Jones JB (2003) Finite mixture distributions, sequential likelihood, and the EM algorithm. *Econometrica* 71(3): 933–946.

Arcidiacono P, Miller RA (2011) Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. *Econometrica* 79(6):1823–1867.

Assmus G, Farley JU, Lehmann DR (1984) How advertising affects sales: Meta-analysis of econometric results. *J. Marketing Res.* 21(1):65–74.

Bajari P, Benkard CL, Levin J (2007) Estimating dynamic models of imperfect competition. *Econometrica* 75(5):1331–1370.

Bangert-Drowns RL, Kulik JA, Kulik C-LC (1991) Effects of frequent classroom testing. *J. Educational Res.* 85(2):89–99.

Basu AK, Lal R, Srinivasan V, Staelin R (1985) Salesforce compensation plans: An agency theoretic perspective. *Marketing Sci.* 4(4):267–291.

Beauchamp A (2012) Regulation, imperfect competition, and the U.S. abortion market. Working paper, Boston College, Chestnut Hill, MA.

Bijmolt THA, Van Heerde HJ, Pieters RGM (2005) New empirical generalizations on the determinants of price elasticity. *J. Marketing Res.* 42(2):141–156.

Chan TY, Li J, Pierce L (2014) Compensation and peer effects in competing sales teams. *Management Sci.* Forthcoming.

Chevalier J, Goolsbee A (2009) Are durable goods consumers forward-looking? Evidence from college textbooks. *Quart. J. Econom.* 124(4):1853–1884.

Copeland A, Monnet C (2009) The welfare effects of incentive schemes. *Rev. Econom. Stud.* 76(1):96–113.

Darmon R (1997) Selecting appropriate sales quota plan structures and quota-setting procedures. *J. Personal Selling Sales Management* 17(1):1–16.

Dubé J-P, Hitsch G, Jindal P (2012) The joint identification of utility and discount functions from stated choice data: An application to durable goods adoption. Working paper, University of Chicago, Chicago.

Efron B (1979) Bootstrap methods: Another look at the jackknife. *Ann. Statist.* 7(1):1–26.

Ferrall C, Shearer B (1999) Incentives and transactions costs within the firm: Estimating an agency model using payroll records. *Rev. Econom. Stud.* 66(2):309–338.

Finger SR (2012) An empirical analysis of R&D competition in the chemicals industry: Incorporating unobserved heterogeneity into a dynamic game estimator. Working paper, University of South Carolina, Columbia.

Frederick S, Loewenstein G, O'Donoghue T (2002) Time discounting and time preference: A critical review. *J. Econom. Literature* 40(2):351–401.

Heath C, Larrick RP, Wu G (1999) Goals as reference points. *Cognitive Psych.* 38(1):79–109.

Heide CP (1999) *Dartnell's 30th Sales Force Compensation Survey*, 30th ed. (Dartnell Corp, Durham, NC).

Holmstrom B (1979) Moral hazard and observability. *Bell J. Econom.* 10(1):74–91.

Holmstrom B, Milgrom P (1987) Aggregation and linearity in the provision of intertemporal incentives. *Econometrica* 55(2): 303–328.

Hotz VJ, Miller RA (1993) Conditional choice probabilities and the estimation of dynamic models. *Rev. Econom. Stud.* 60(3): 497–529.

Jiang R, Palmatier RW (2009) Structural estimation of a moral hazard model: An application to business selling. Working paper, University of California, Davis, Davis.

Joseph K, Kalwani M (1998) The role of bonus pay in sales force compensation plans. *Indust. Marketing Management* 27(2): 147–159.

Judd KL (1998) *Numerical Methods in Economics* (MIT Press, Cambridge, MA).

Kim SK (1997) Limited liability and bonus contracts. *J. Econom. Management Strategy* 6(4):899–913.

Lal R, Srinivasan V (1993) Compensation plans for single- and multi-product sales forces: An application of the Holmstrom-Milgrom model. *Management Sci.* 39(7):777–793.

Magnac T, Thesmar D (2002) Identifying dynamic discrete choice models. *Econometrica* 70(2):801–816.

Mirrlees JA (1999) The theory of moral hazard and unobservable behaviour: Part I. *Rev. Econom. Stud.* 66(1):3–21. [Mimeo, Nuffield College, Oxford, UK, 1975.]

Misra S, Nair H (2011) A structural model of sales-force compensation dynamics: Estimation and field implementation. *Quant. Marketing Econom.* 9(3):211–257.

Oyer P (1998) Fiscal year ends and nonlinear incentive contracts: The effect on business seasonality. *Quart. J. Econom.* 113(1): 149–185.

Oyer P (2000) A theory of sales quotas with limited liability and rent sharing. *J. Labor Econom.* 18(3):405–426.

Paarsch HJ, Shearer B (2000) Piece rates, fixed wages, and incentive effects: Statistical evidence from payroll records. *Internat. Econom. Rev.* 41(1):59–92.

Park E-S (1995) Incentive contracting under limited liability. *J. Econom. Management Strategy* 4(3):477–490.

Raju JS, Srinivasan V (1996) Quota-based compensation plans for multi-territory heterogeneous salesforces. *Management Sci.* 42(10):1454–1462.

Rao RC (1990) Compensating heterogeneous sales-forces: Some explicit solutions. *Marketing Sci.* 9(4):319–341.

Rust J (1987) Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica* 55(5):999–1033.

Rust J (1994) Structural estimation of Markov decision processes. Engle RE, McFadden D, eds. *Handbook of Econometrics*, Vol. 4 (Elsevier-North Holland, Amsterdam), 3082–3139.

Shen Q (2013) A dynamic model of entry and exit in a growing industry. Working paper, University of Pennsylvania, Philadelphia.

Steenburgh T (2008) Effort or timing: The effect of lump-sum bonuses. *Quant. Marketing Econom.* 6(3):235–256.

Sudhir K (2001) Structural analysis of manufacturer pricing in the presence of strategic retailer. *Marketing Sci.* 20(3):244–264.

Training (2008) Quota-based compensation on the rise. (October 15) http://www.trainingmag.com/article/quota-based-compensation-rise.

Yao S, Mela CF, Chiang J, Chen Y (2012) Determining consumers' discount rates with field studies. *J. Marketing Res.* 49(6): 822–841.

Zoltners AA, Sinha P, Lorimer SE (2008) Sales force effectiveness: A framework for researchers and practitioners. *J. Personal Selling Sales Management* 28(2):115–131.