Homogenous Contracts for Heterogeneous Agents: Aligning Salesforce Composition and Compensation*

Øystein Daljord Sanjog Misra Harikesh S. Nair

Past versions: June 2012; Feb 2013, May 2015. This version: Sept 6, 2015

Abstract

Observed contracts in the real-world are often very simple, partly reflecting the constraints faced by contracting firms in making the contracts more complex. We focus on one such rigidity, the constraints faced by firms in fine-tuning contracts to the full distribution of heterogeneity of its employees. We explore the implication of these restrictions for the provision of incentives within the firm. Our application is to salesforce compensation, in which a firm maintains a salesforce to market its products. Consistent with ubiquitous real-world business practice, we assume the firm is restricted to fully or partially set uniform commissions across its agent pool. We show this implies an interaction between the composition of agent types in the contract and the compensation policy used to motivate them, leading to a "contractual externality" in the firm and generating gains to sorting. This paper explains how this contractual externality arises, discusses a practical approach to endogenize agents and incentives at a firm in its presence, and presents an empirical application to salesforce compensation contracts at a US Fortune 500 company that explores these considerations and assesses the gains from a salesforce architecture that sorts agents into divisions to balance firm-wide incentives. Empirically, we find the restriction to homogenous plans significantly reduces the payoffs of the firm relative to a fully heterogeneous plan when it is unable to optimize the composition of its agents. However, the firm's payoffs come very close to that of the fully heterogeneous plan when it can optimize both composition and compensation. Thus, in our empirical setting, the ability to choose agents mitigates the loss in incentives from the restriction to uniform contracts. We conjecture this may hold more broadly.

* *Daljord*: Booth School of Business, University of Chicago, daljord@chicagobooth.edu; *Misra*: Booth School of Business, University of Chicago, sanjog.misra@chicagobooth.edu; *Nair*: Graduate School of Business, Stanford University, harikesh.nair@stanford.edu. We thank Guy Arie, Nick Bloom, Francine Lafontaine, Ed Lazear, Sridhar Moorthy, Paul Oyer, Michael Raith, Kathryn Shaw, Chuck Weinberg, Jeff Zwiebel; our discussant at the QME conference, Curtis Taylor; and Lanier Benkard and Ed Lazear in particular for their useful comments and suggestions. We also thank seminar participants at Arizona-Economics, Harvard-Economics, Michigan-Economics, MIT-Sloan, Stanford-GSB, UBC-Saunder, UC-Davis-GSM, UT-Austin-McCoombs, UToronto-Rotman, and at the 2012 IOFest, Marketing Dynamics, World Congress of Econometrics and QME conferences for useful feedback. The usual disclaimer applies.

1 Introduction

In many interesting market contexts, firms face rigidities or constraints in fine-tuning their contracts to reflect the full distribution of heterogeneity of the agents they are contracting with. For example, auto-insurance companies are often prevented by regulation from conditioning their premiums on consumer characteristics like race and credit scores. Royalty rates in business-format franchising in the US are typically constrained by norm to be the same across all franchisees in a given chain.[1] Wholesales contracts between manufacturers and downstream retailers in the US typically involve similar wholesales prices to all downstream retailers within a given geographic area due to Robinson-Patman considerations. In salesforces, the context of the empirical example in this paper, incentive or commission rates on output are invariably set the same across all sales-agents within a firm. For instance, a firm choosing a salary + commission salesforce compensation scheme typically sets the same commission rate for every sales-agent on its payroll, in spite of the fact that exploring the heterogeneity and setting an agent-specific commission may create theoretically better incentives at the individual level. While reasons are varied, full or partial uniformity of this sort is well documented to be an ubiquitous feature of real-world salesforce compensation (Rao 1990; Mantrala et al. 1994; Raju and Srinivasan 1996; Zoltners et al. 2001; Lo et al. 2011).

The focus of this paper is on the implications to the principal of this restriction to similar contract terms across agents. We do not take a strong stance on the source of the uniformity, but focus on the fact that any such uniformity in the contract implies that agents and contract terms have to be chosen jointly. In the salesforce context, for example, this creates an interaction between the composition of agent types in the contract and the compensation policy used to motivate them, leading to a "contractual externality" in the firm and generating gains to sorting. This paper explains how this contractual externality arises, discusses a practical approach to endogenize agents and incentives at a firm in its presence, and presents an empirical application to salesforce compensation

[1] Quoting Lafontaine and Blair (2009, pp. 395-396), "Economic theory suggests that franchisors should tailor their franchise contract terms for each unit and franchisee in a chain. In practice, however, contracts are *remarkably uniform* across franchisees at a point in time within chains [emphasis ours]..[]..a business-format franchisor most often uses a single business-format franchising contract—a single royalty rate and franchise fee combination— for all of its franchised operations that join the chain at a given point..[]..Thus, uniformity, especially for monetary terms, is the norm."

contracts at a US Fortune 500 company that explores these considerations and assesses the gains from a salesforce architecture that sorts agents into divisions to balance firm-wide incentives.

As a motivating example, consider a firm that has chosen a salary + commission scheme. Suppose all agents have the same productivity, but there is heterogeneity in risk aversion amongst the agents. The risk averse agents prefer that more of their pay arises from fixed salary, but the less risk averse prefer more commissions. When commissions are restricted to be uniform across agents, including the more risk averse types in the firm implies the firm cannot offer high commissions. Dropping the bottom tail of agents from the firm may then enable the firm to profitably raise commissions for the rest of agents. Knowing that, the firm should choose agents and commissions jointly. This is our first point: the restriction to uniformity implies the composition and compensation are co-dependent. To address this, the contracting problem has to be enlarged to allow the principal to choose the distribution of types in his firm along with the optimal contract form given that type distribution.

Our second point is that uniformity implies the presence of a sales-agent in the firm imposes an externality on the other agents in the pool through its effect on the shape of the common element of the incentive contract. For instance, suppose agents are homogenous in all respects except their risk aversion, and there are three agents, A, B and C who could be employed, with C the most risk averse. C needs more insurance than A and B and retaining him requires a lower common commission rate. It could then be that A and B are worse off with C in the firm (lower commissions). Thus, the presence of the low-type agent imposes an externality on the other sales-agent in the firm through the endogeneity of contract choice.

This externality can be substantial when agent types are *multidimensional* (for example, when agents are heterogeneous in risk aversion, productivity and costs of expending sales effort). Consider the above example. It may be optimal for the principal to rank the three agents on the basis of their risk aversion and to drop the "low-type" C from the agent pool. Hence, if risk aversion were the only source of heterogeneity, and we enlarge the contracting problem to allow the principal to choose both the optimal composition and compensation, it may be that "low-types" like C impose little

externalities because they are endogenously dropped from the firm. Now, consider what happens when types are multidimensional. Suppose in addition to risk aversion, agents are heterogeneous in their productivity (in the sense of converting effort into output), and C, the most risk averse is the most productive. Then, the principal faces a tradeoff: dropping C from the pool enables him to set more high powered incentives to A and B, but also entails a large loss in output because C is the most productive. In this tradeoff, it may well be that the optimal strategy for the principal is to retain C in the agent-pool and to offer all the lower common commission induced by his presence. Thus multidimensional types increase the chance that the externalities we discussed above persist in the optimally chosen contract. More generally, multidimensionality of the type space also points to the need for a theory to describe who should be retained and who should be let go from the salesforce, because agents cannot be ranked as desirable or undesirable on the basis of any one single variable.

Our main question explores the co-dependence between composition and compensation. We ask to what extent composition and compensation complement each other in realistic salesforce settings. We use an agency theoretic set-up in which the principal chooses both the set of agents to retain in the firm and the optimal contract to incentivize the retained agents. We use our model to simulate how the contract form changes when the distribution of ability changes, which helps measure the size of externalities and assess the value of policies to mitigate it. A realistic assessment of these issues is dependent on the distribution of heterogeneity in the agent pool, and hence is inherently an empirical question.

We leverage access to a rich dataset containing the joint distribution of output and contracts for all sales-agents at a Fortune 500 contact lens company in the US. We build on our analysis in Misra and Nair (2011), which used these data to identify primitive agent parameters (cost of effort, risk aversion, productivity), and to estimate the multidimensional distribution of heterogeneity in these parameters across agents at the firm. In the data, agents are paid according to a nonlinear, quarterly incentive plan consisting of a salary and a linear commission which is earned if realized sales are above a contracted quota and below a pre-specified ceiling. The nonlinearity of the incentive

contract creates dynamics in the agent's actions, causing the agent to optimally vary his effort profile as he moves closer to or above his quota. The joint distribution of output and the distance to the quota thus identify "hidden" effort in this moral hazard setting. In Misra and Nair (2011), this identification strategy is incorporated into a structural model of agent's optimization behavior to recover the primitives underpinning agent types. Here, we use these estimates as an input into a model of simultaneous contact form and agent composition choice for the principal.

Solving this model involves computing a large-scale combinatorial optimization problem in which the firm chooses one of 2^N possible salesforce configurations from amongst a pool of N potential agents, and solving for the optimal common incentive contract for the chosen pool. To find the optimal composition of agents, we can always enumerate the profits for all 2^N combinations when N is small. If N is large, simple enumeration algorithm is practically infeasible (N is around 60 in our focal firm, and could number in the hundreds or thousands in other applications). Since agents are allowed a multidimensional type space, it is generally not possible to find a simple cut-off rule where agents above some parameter threshold are retained. Exploiting a characterization of the optimal solution for a class of composition-compensation problems, we derive an algorithm that allows us to search the exploding composition-compensation space by reducing it to a standard optimization program for continuous functions on compact sets. The reduction makes the execution time of the algorithm independent of the composition space itself. In examples we compute, we can search a space of 2^{5000} composition-compensations in fractions of a second. A power-set search of a space of that size is otherwise prohibitive.

We use the algorithm to simulate counterfactual contracts and agent pools at the estimated parameters. We explore to what extent a change in composition of the agents affects the nature of optimal compensation for those agents, and quantify the profit impact of jointly optimizing over composition and compensation. We find that allowing the firm to optimize the composition of its types has bite in our empirical setting. When the firm is restricted to homogenous contracts and no optimization over types, we estimate its payoffs are significantly lower than that under fully heterogeneous contracts. However, the payoffs under homogeneous contracts when the firm can

optimize both composition and compensation come very close to that under fully heterogeneous contracts. We demonstrate these results are robust to parameter uncertainty the principal may have about its estimates of the agent's types. Overall, we find the ability to choose agents helps balance significantly the loss in incentives from the restriction to homogeneity. We conjecture this may be broadly relevant in other settings and may help rationalize the prevalence of homogenous contracts in many salesforce settings in spite of the profit consequences of reduced incentives.

We then simulate a variety of salesforce architectures in which the firm sorts its salesagents into divisions. We restrict each division to offer a uniform commission to all within its purview, but allow commissions to vary across divisions. We then simultaneously solve for the optimal commissions and the optimal allocation of agents to divisions. In the context of our empirical example, we find that a small number of divisions generates profits to the principal that come very close to that under fully heterogeneous contracts. If the firm is allowed to choose its composition as well, this profit gain is achieved with even fewer divisions. The main take-away is that simple contracts combined with the ability to choose agents seem to do remarkably well compared to more complex contracts, at least in the context of our empirical example.

Our analysis is related to a literature that emphasizes the "selection" effect of incentives, for example, Lazear (2000a)'s famous analysis of Safelite Glass Corporation's incentive plan for windshield installers, in which he demonstrates that higher-ability agents remain with the company after it switched from a straight salary to a piece-rate; or Bandiera et al. (2007)'s analysis of managers at a fruit-picking company, who started hiring more high ability workers after they were switched to a contract in which pay depends on the performance of those workers. Lazear and Bandiera et al. present models of how the types of agents that sort into or are retained at the firm changes in response to an *exogenously* specified piece-rate. In our set-up, the piece-rate *itself* changes as the set of agents at the firm changes, because the firm jointly chooses the contract and the agents. The *endogenous* adjustment of the contract as the types change is key to our story. The closest we know to our point in the literature is Lazear (2000b), who shows that firms may choose incentives to attract agents of high ability. Unlike our set-up though, Lazear considers unidimensional agents in

an environment with no asymmetric information or uncertainty, so many of the contractual forces identified here are not a feature of his analysis. Lazear also makes a broader point that piece-rates have an advantage of helping manage heterogeneity within the firm. Our analysis here has parallels to this insight.

The literature on relative performance schemes and on teams (Holmstrom 1982; Kandel and Lazear 1992; Hamilton et al. 2003; Misra, Pinker and Shumsky 2004) has identified other contexts wherein one agent's characteristics or actions substantively affects another's welfare via an interaction with incentives. The contractual externality we identify persists when agents have exclusive territories and there are no across-agent complementarity or substitution effects in output, and is relevant even when contracts are absolute and not relative. It is thus distinct from the mechanisms identified in this literature. A related literature on contract design in which one principal contracts with many agents focuses on the conditions where relative incentive schemes arise endogenously as optimal, and not on the question of the joint choice of agents and incentives, which is our focus here. Broadly speaking, the relative incentive scheme literature focuses on the value of contracts in filtering out common shocks to demand and output, and on the advantages of contracting on the ordinal aspect of outputs when output is hard to measure (e.g, Lazear and Rosen 1981; Green and Stokey 1983; Mookerjee 1984; Kalra and Shi 2001; Lim et al. 2009; Ridlon and Shin 2010). Common shocks and noise in the output measure are not compelling features of our empirical setting which involves selling of contact-lenses to optometricians, for which seasonality and co-movement in demand is limited, and sales (output) are precisely tracked. A small theoretical literature also emphasizes why a principal may choose a particular type of agent in order to signal commitment to a given policy (e.g., shareholders may choose a "visionary" CEO with a reputation for change-management so as to commit to implementing change within the firm: e.g., Rotemberg and Saloner, 2000). Our point, that the principal may choose agents for incentive reasons, is distinct from that in this literature which focuses on commitment as the rationale of the principal for its choice of agents. A related theoretical literature has also noted that contracts may signal information that affects the set of potential employees or franchisees a principal may contract with (Desai and Srinivasan 1995; Godes

and Mayzlin 2012), without focusing on the principal's choice of agents explicitly.

Findings related to our results here − that it may be optimal for the principal to drop some agents and to group together agents of differing types into divisions in order to achieve appropriate separation − are also reflected in a small theoretical literature on *multidimensional* screening, canonical examples of which are discussed in the context of nonlinear pricing in Armstrong (1996) and Rochet and Chone (1998). For instance, Armstrong (1996) shows that the optimal price schedule for a multiproduct firm facing consumers with (unknown) multidimensional types may involve excluding some consumers from its products in order to extract more revenue from the high value consumers. Rochet and Chone (1998) show such optimal contracts may typically involve some degree of "bunching", so that consumers of different types choose the same bundle of products. While there are these parallels, note this literature focuses on adverse selection as the manifestation of asymmetric information. In contrast, ours is a problem with moral hazard, which has a more complicated structure because the unobservable "type" of the agent (i.e., hidden effort) changes with the contract.

Our model predicts that agents and incentives across firms are simultaneously determined and has implications for two related streams of empirical work. One stream measures the effect of incentives on workers, and tests implications of contract theory using data on observed contracts and agent characteristics across firms (see Pendergast 1999 for a review). In an important contribution to the econometrics in this area, Ackerberg and Botticini (2002) note that when agents are endogenously matched to contracts, the correlation observed in data between outcomes and contract characteristics should be interpreted with caution. A potential for confounds arises from unobserved agent characteristics that may potentially be correlated with both outcomes and contract forms. The resulting omitted variables problem may result in endogeneity biases when trying to measure the causal effect of contracts on outcomes. Our model, which provides a rationale for why agent and contracts characteristics are co-determined across firms, has similar implications for empirical work using across-agent data. The model implies that the variation in contract terms across firms is endogenous to worker characteristics at those firms. While Ackerberg and Botticini (2002) stress

the omitted variables problem, the endogeneity implied by our model derives from the *simultaneity* of contracts and agents.

A second stream pertains to work that has measured complementarities in human resource practices within firms, testing the theory developed in Milgrom and Roberts (1990) and Holmstrom and Milgrom (1994), amongst others. This theory postulates that human resource activities like worker training and incentive provision are complementary activities. A large body of empirical work has measured the extent of these complementarities using across-firm data correlating worker productivity with the incidence of these activities (e.g., Ichniowski, Shaw and Prennushi 1997 and others). Our model predicts that workers and incentives (or HR practices, more generally) are optimally jointly chosen. When better workers also have corresponding better productivity, the simultaneity of worker choice and HR practices implies the incidence of HR practices are endogenous in productivity regressions, which confounds the measurement of such complementaries using across-firm data. A related implication of our model is that endogenously adjusting common clauses of firm-wide contracts can generate across-agent dependencies in output that generate *indirect* complementarities. If these are not accounted for, it may confound measurement of other sources of *direct* complementarities like peer effects that researchers are interested in measuring using within-firm personnel data. More research and better data are required to address these kinds of difficult econometric concerns in empirical work. We now discuss our model set-up and present the rest of the analysis.

2 Model

A firm wishes to optimize the composition and compensation of its salesforce. The firm is assumed to know all the agent's relevant characteristics with certainty, but is not able to observe their effort with certainty. Conditional on the group of agents, the problem is similar to the classic hidden action problem of Holmstrom (1979). Principal certainty of the agents characteristics may be a reasonable assumption for the retention problem where the firm has known the agents for a long time (like in our application), but is far more questionable as a point of departure for hiring. Attention is

therefore here restricted to the retention problem, on who the firm should retain when there are contractual externalities that depend on the composition. To simplify the application, we abstract away from uncertainty about the agents types to avoid issues of learning and adverse selection. Learning about agent type is not of first-order importance in our application because most agents have been with the firm for a long time (mean tenure 9 years). However, this may be an important dynamic for new workers. We discuss these issues in more detail later in the paper.

Reflecting the empirical application, the firm has divided its potential market into N geographic territories, and the maximum demand at the firm is for N sales-agents.[2] There are N heterogeneous sales-agents indexed by $i = 1...N$ currently employed or employable. Let $\mathbb{M}_N$ denote the power-set spanned by N (that is, all possible sub-salesforces that could be generated by N), and $\mathbb{W}_{\mathcal{M}}$ the set of compensation contracts possible for a specific sub-salesforce $\mathcal{M}$. Let S_i denote agent i's output, $\mathcal{W}(S_i)$ his wages conditional on output, and $\mathcal{F}(S_i|e_i)$ denote the CDF of output conditional on effort choice, e_i. Effort e_i is privately observed by the agent and not by the principal, while output S_i is observed by both the agent and the principal, and hence is contractible. As is common in the agency literature, we assume that the agent chooses effort before sales are realized, that both he and the principal share the same beliefs about the conditional distribution of output ($\mathcal{F}(S_i|e_i)$) (common knowledge about outcomes). Since sales are stochastic, the principal cannot back out the hidden effort from realized output, which generates the standard moral hazard problem.

The principal maximizes,

$$\max_{\mathcal{M}\in\mathbb{M}_N, \mathcal{W}\in\mathbb{W}_{\mathcal{M}}} \Pi = \int \sum_{i\in\mathcal{M}} \left[S_i - \mathcal{W}(S_i)\right] d\mathcal{F}(S_i|e_i) \tag{1}$$

where the control, $(\mathcal{M}, \mathcal{W})$, is the set of active agents and their compensation. The maximization is subject to the Incentive Compatibility (IC) constraints, that the effort chosen by each agent i is optimal,

$$e_i = \arg\max_e \int U\left(\mathcal{W}(S_i), C(e;\mu_i)\right) d\mathcal{F}(S_i|e) \quad \forall i \in \mathcal{M} \tag{2}$$

[2] More generally, the need for a maximum of N agents can be thought of as implying that total profit for the firm is concave in N.

and the Individual Rationality (IR) constraints that each active agent i receives at least expected reservation utility $\widetilde{U}_i^0$ from staying with the firm and working under the suggested contract,

$$\int U\left(\mathcal{W}\left(S_i\right), C\left(e;\mu_i\right)\right) d\mathcal{F}\left(S_i|e_i\right) \geq \widetilde{U}_i^0 \quad \forall i \in \mathcal{M} \tag{3}$$

The above set-up endogenizes the principal's choice of the agent pool in the following way. The principal knows each agent's type (including reservation utility). He designs a contract such that the IR constraints in equation (3) are satisfied only for the set of agents in $\mathcal{M}$ and violated for all others. This contract provides the chosen set of agents in $\mathcal{M}$ enough utility to stay; the rest are better off pursuing their outside option. Thus the contract endogenously induces the preferred agents to stay and the others to quit.[3] To complete the model, we also need to specify what happens to demand from a territory managed by an agent if that agent leaves. We assume that sales equivalent to the intercept in the output function (discussed below) continue to accrue to the firm even if no agent operates in that territory. This encapsulates the notion that a base level of sales will be generated to the firm even in the absence of any marketing or salesforce effort. Given this, an equivalent interpretation of the principal's decision to offer a contract that induces an agent to quit is that he has decided to vacate the territory managed by the agent. Obtaining only the base level of sales from the territory but offering an improved contract to the others, is more beneficial than retaining the agent and incurring the added pay and contractual externalities induced by his presence.

Equivalent Bi-level Setup We can reformulate the problem by allowing the principal to choose the optimal contract in a first step, and then solving point wise for the optimal configuration for the chosen contract. The program described above is equivalent to the case where the principal maximizes,

$$\Pi = \max_{\mathcal{W}\in\mathbb{W}_{\mathcal{M}}} \int \sum_{i\in\mathcal{M}_{\mathcal{W}}} \left[S_i - \mathcal{W}\left(S_i\right)\right] d\mathcal{F}\left(S_i|e_i\right)$$

[3]This need not be implemented by explicitly "firing" an agent. Not offering raises as part of a restructuring exercise, or providing only reduced pay could induce the outcome.

with,

$$\mathcal{M}_{\mathcal{W}} = \underset{\mathcal{M}\in\mathbb{M}_N}{\arg\max} \int \sum_{i\in\mathcal{M}} \left[S_i - \mathcal{W}\left(S_i\right)\right] d\mathcal{F}\left(S_i|e_i\right)$$

subject to the IC and IR constraints as before. Contractual externalities arise because some elements of the contract $\mathcal{W}\left(S_i\right)$ are common across agents, which makes the problem non-separable across agents. Since $\mathcal{M}_{\mathcal{W}} \in \mathbb{M}_N$ is point-wise the optimal sub-salesforce plan for each considered contract $\mathcal{W}$, the solution to this revised problem returns the solution to the original program. Representing the program this way helps understand our numerical algorithm for solution more clearly.

3 Application Setting

To illustrate the main forces at work clearly and to operationalize the setup above for our empirical setting, we now discuss the parametric assumptions we impose. We employ a version of the well-known Holmstrom and Milgrom (1987) model for two reasons:

1. The model has a closed-form solution that is useful from both an illustrational and a computational point of view.
2. The optimal contracts are linear (salary plus commission) which is empirically relevant.

Though linear contracts are used as the illustrational vehicle, the qualitative aspects of the setup holds more generally for any multilateral contracting problem with inter-agent externalities induced by common contractual components. Each agent i is described completely by a tuple $\{h_i, k_i, d_i, r_i, \sigma_i, U_i^o\}$. The elements of the tuple will become clear in what follows. Sales are assumed to be generated by the following functional,

$$S_i = h_i + k_i e_i + \sigma_i \varepsilon_i \tag{4}$$

This functional has been used in the literature (see e.g. Lal and Srinivasan 1992) and interprets h as the expected sales in the absence of selling effort (i.e. $\mathbb{E}\left[S_i|e_i = 0\right] = h_i$), k_i as the marginal productivity of effort and σ_i^2 as the uncertainty in the sales production process. As is usual, we assume that the firm only observes S_i and knows $\{h_i, k_i, \sigma_i\}$ for all agents. The density $\mathcal{F}\left(S_i|e_i\right)$ is induced by the density of ε_i. Under linear contracts, compensation is $\mathcal{W}\left(S_i\right) = \alpha_i + \beta S_i$, where

α_i is the salary, which can be agent-specific, and β is the commission rate which is common across agents. The agent is assumed to have a CARA utility function $U_i(W_i) = -\exp\{-r_i W_i\}$, defined over wealth W_i, which in turn is linear in output and quadratic (convex) in the cost of effort, i.e. $W_i = \alpha_i + \beta S_i - \frac{d_i}{2} e_i^2$. The agent chooses effort to maximize expected utility, where the expectation is taken over the shocks to sales: $\mathbb{E}_\epsilon[U_i(W_i)] = -\int \exp\left[-r_i\left(\alpha_i + \beta S_i - \frac{d_i}{2} e_i^2\right)\right] d\mathcal{F}(\varepsilon_i)$. The implied Certainty Equivalent for the agent is,

$$\mathcal{CE}_i = \alpha_i + \beta(h_i + k_i e_i) - \frac{d_i}{2} e_i^2 - \frac{r_i}{2}\beta^2\sigma_i^2 \tag{5}$$

maximizing which implies the optimal effort choice for the agent is, $e_i(\beta) = \beta\frac{k_i}{d_i}$.

3.1 The Principal's Problem

The principal treats agents as exchangeable and cares only about expected profits, which he maximizes subject to the (IC, IR) constraints to find the agent-specific salaries and common commission rate (α_i, β),

$$\begin{aligned} \max_{(\alpha_i,\beta)} \mathbb{E}[\Pi] = \mathbb{E}\sum_{i=1}^{N}(S_i - \beta S_i - \alpha_i) \quad \text{s.t.} \\ IC:\ e_i(\beta) = \beta\frac{k_i}{d_i} \qquad \forall i = 1.,,.N \\ IR:\ \mathcal{CE}_i \geq U_i^o \qquad \forall i = 1.,,.N \end{aligned}$$

In the above, U_i^0 is the certainty equivalent of the agent's outside option utility. The principal's problem can be simplified by incorporating the IC constraint,

$$\mathbb{E}[\Pi] = \sum_{i=1}^{N}(\mathbb{E}(S_i) - \beta\mathbb{E}(S_i) - \alpha_i) = \sum_{i=1}^{N}(1-\beta)(h_i + k_i e_i(\beta)) - \alpha_i \tag{6}$$

Further if the IR constraint is binding we can write,

$$\alpha_i(\beta) = U_i^o - [\beta(h_i + k_i e_i) - \frac{d_i}{2} e_i^2 - \frac{r_i}{2}\beta^2\sigma_i^2] \tag{7}$$

and substituting in we have,

$$\begin{aligned}\mathbb{E}[\Pi] &= \sum_{i=1}^{N}[\ (1-\beta)(h_i + k_i e_i(\beta)) - U_i^o + \left\{\beta(h_i + k_i e_i(\beta)) - \tfrac{d_i}{2} e_i(\beta)^2 - \tfrac{r_i}{2}\beta^2\sigma_i^2\right\}\ ] \\ &= \sum_{i=1}^{N}[h_i + k_i e_i(\beta) - U_i^o - \frac{d_i}{2} e_i(\beta)^2 - \frac{r_i}{2}\beta^2\sigma_i^2] \end{aligned} \tag{8}$$

Differentiating with respect to β we get,

$$\frac{\partial \mathbb{E}[\Pi]}{\partial \beta} = \sum_{i=1}^{N}[k_i e_i'(\beta) - d_i e_i'(\beta) - r_i\beta\sigma_i^2] \tag{9}$$

where, $e_i'(\beta) = \frac{\partial e_i(\beta)}{\partial \beta} = \frac{k_i}{d_i}$. Solving (9) gives the optimal common commission,

$$\beta = \frac{1}{1+\gamma} \tag{10}$$

where, $\gamma = \frac{\sum_{i=1}^{N} r_i\sigma_i^2}{\sum_{i=1}^{N} \frac{k_i^2}{d_i}}$, is an aggregate measure of the distribution of types within the firm. The salary, α_i can be obtained by substitution.

Equation (10) encapsulates the effect of each agent type on the contract: the optimal α_i, β depends on the distribution of characteristics of the entire agent pool. Equation (10) demonstrates the *contractual externality* implied by homogeneity restrictions: when an agent joins or leaves the salesforce, he affects everyone else by changing the optimal β. Equation (10) also helps us build intuition about how the distribution of types in the firms shifts the common commission rate. Holding everything else fixed, an increase in the level of risk aversion in the salesforce increases γ, which reduces the commission rate as expected. As normalized productivity, $\frac{k_i^2}{d_i}$, increases in the salesforce, γ falls, and optimal commissions increase as expected. Both are intuitive. An assessment of the net effect on commissions as *both* risk aversion and normalized productivities change is more difficult, as it depends on the extent of affiliation between these parameters in the salesforce (e.g., whether more risk-averse agents are more productive or less).

To see how the profit function depends overall on agent's types, we can write equation (8)

evaluated at the optimal commission β as,

$$\mathbb{E}\left[\Pi\left(\beta\right)\right] \quad = \quad \sum_{i=1}^{N}\left(h_i - U_i^o\right) + \frac{\beta}{2}\left(\sum_{i=1}^{N}\frac{k_i^2}{d_i}\right) \tag{11}$$

Noting that d_i is the agent's cost of effort, we can think of $1/d_i$ as a measure of the agent's efficiency − those with higher $1/d_i$ expend the same effort at lesser cost. Equation (11) can be interpreted as decomposing the firm's total profits at the optimally chosen incentive level into two components. The first comprises the total baseline revenue from each agent when each is employed, but expending zero effort, $(h_i - U_i^0)$. The second is the optimal commission rate times a weighted average of each agent's efficiency $(1/d_i)$, where the weights correspond to each agent's productivity (k_i^2). The first part is the insurance component of the incentive scheme, while the second part reflects incentives. Equation (11) also shows that profits are separable across agents except for the choice of β. In the absence of endogenizing β, the decision to retain an agent i in the pool has no bearing on the decision to retain another.

Substituting for the optimal β from equation (10), we can write the total payoff to the principal with optimally chosen incentives as,

$$\mathbb{E}\left[\Pi\right] = \sum_{i=1}^{N}\left(h_i - U_i^o\right) + \frac{1}{2}\frac{\left(\sum_{i=1}^{N}\frac{k_i^2}{d_i}\right)^2}{\left(\sum_{i=1}^{N}\frac{k_i^2}{d_i} + \sum_{i=1}^{N} r_i\sigma_i^2\right)} \tag{12}$$

Equation (12) shows that at the optimal β^*, the payoff across agents is no longer separable across types. Equation (12) defines the firm's optimization problem over the N agent types given optimal choice of incentives for each sub-configuration.

To build intuition, suppose the firm has the option of retaining three agents with low, medium and high risk aversion. If forced to retain all three on the payroll on a salary + commission scheme, the firm is not able to have a very high-powered incentive scheme with a high commission rate because the high risk-averse agent has to be provided significant insurance. Firing the high risk averse agent will allow the firm to optimally charge a higher commission rate to the remaining two agents. Depending on the productivity and cost parameters of the three agents, we can construct examples where the payoffs to the firm with two agents and the high commission are higher than

with three agents and the lower commission.

A Simple 3-Agent Example

In the three agent example below, it is shown that the principal may not find the most prolific sales agent attractive if that agent sufficiently skews the incentives of the other agents. The parameters are set to $(h, k, \sigma) = 1$ for all agents. The remaining parameters are given by:

Parameter	Agent 1	Agent 2	Agent 3
d	$\frac{1}{5}$	$\frac{3}{2}$	$\frac{1}{2}$
r	4	3	5
U^0	$\frac{9}{4}$	1	0

Agent 1 has the lowest cost of effort d, agent 2 has the lowest risk aversion and agent 3 has the lowest outside utility U^0. Because of the multidimensional type space, it is not immediately clear which agents are the most profitable.

The first three rows in the Table below gives the individually optimal contracts for all three agents. With no contractual externalities, the firm would retain all agents and provide each agent a commission tailored to his type and set a salary that just leaves each his reservation utility.

	Composition	Profits	Commission Rate	Sales	Compensation
Uniform commissions	$\mathcal{M}$	$\mathbb{E}[\Pi(\beta)]$	β	$\mathbb{E}[\sum S_i]$	$\mathbb{E}[\mathcal{W}_{\mathcal{M}}]$
	$(1,0,0)$	0.14	0.56	3.78	3.64
1 agent	$(0,1,0)$	0.06	0.18	1.12	1.06
	$(0,0,1)$	1.29	0.29	1.57	0.29
	$(1,1,0)$	0.02	0.45	4.54	4.52
2 agents	$(1,0,1)$	1.28	0.44	5.06	3.78
	(0,1,1)	**1.33**	**0.25**	**2.67**	**1.33**
3 agents	$(1,1,1)$	1.24	0.39	5.99	4.74
Individually optimal contracts	$(1,1,1)$	1.49	(0.56,0.18,0.29)	6.47	4.99

In the fully heterogenous plan summarized in the last row, there are no externalities. Agent 1 gets the largest commission (56%), agent 2 the smallest (18%), and the firm makes a profit of 1.49. Now consider what happens when the firm is restricted to a common commission, but individual salaries for each agent. The results are given in the upper rows. Solving for each configuration, we find that the firm would optimally drop agent 1 from the pool (expected payoff of 1.33 with an optimal common commission rate to agents 2 and 3 of 25%). Including agent 1 in the composition

requires the firm to set a higher commission rate (39%), which is too high for the other agents, reducing the firm's payoffs to 1.24.

Looking at the first two rows of the two agent compositions, agent 1's presence in the pool is seen to exert skew the incentives of agent 2 and 3 away from their individually optimal commissions. If only agent 1 is retained, he would be paid a high-powered commission rate of 56%, while agents 2 and 3 prefer commissions of only 18% and 29% respectively. In this example, the firm is better off dropping the high-powered sales-agent from the pool. Without agent 1, the firm can set an intermediate level of commissions (25%) that provides 2 and 3 better incentives. For a more dramatic effect, consider agent 1 and 2 individually. At the heterogenous contracts at individual commissions 56% and 18%, they bring in profits of 0.20. Under a uniform commission of 45%, the profit is reduced to 0.02, as agent 1 gets underpowered incentives and agent 2 gets overpowered incentives.

The firm makes lower sales with only agents 2 and 3 than with 3 agents (2.67 versus 5.99), but compensation payout is also lower, and the net effect is a higher profit. By changing parameters, we can generate other examples where the low-commission agent is dropped from the pool in order to set high-powered incentives for the remaining agents, which is the mirror-image to this setting. The example below also illustrates the complication induced by multidimensional types: the desirability of an agent cannot be ordered on any one dimension by a simple cut-off rule.

Note that the profits under the optimal uniform commission plan (1.33) come close to that of the individual optimal plan (1.49), despite the restriction to common commissions and the heterogeneity in individually optimal commissions. The example illustrates how careful choice of composition can compensate for the reduced incentives implied by the common contract terms, a point that also shows up in our empirical application below.

The simple example shows how composition and compensation interact in influencing firm profits. While the example has only three agents, it provides a glimpse into the workings of this interaction, and in particular shows that commonly used performance measures such as sales (or even profit contribution) may not be useful in determining which agents should stay. Indeed, in the above example, agent 1 had the highest productivity in terms of sales (3.78 vs. 1.12 and 1.57

for agents 2 and 3) and the second highest profit contribution (.14 vs .06 and 1.29 for agents 2 and 3). However, keeping the agent in the pool distorted the incentives to the others. We conjecture that similar patterns apply more generally in real world sales-forces. To examine this conjecture we use data from a real salesforce below. Before doing so, we introduce our algorithm to compute the optimal salesforce composition.

4 The Solution Algorithm

The composition-compensation problem above is a mixed integer program: the commission is a continuous variable, and the composition choice is a binary integer program. The combinatorial nature of integer programs generally requires computationally demanding algorithms. The example above used a complete enumeration of possible salesforce compositions to examine the profitability of the possible compositions. Since the space of possible compositions grows exponentially in the number of agents, complete enumeration is not feasible for applications to real-world sales forces where there may be hundreds or thousands agents. We provide an algorithm that collapses the mixed integer program to a standard optimization problem of searching a continuous function on a compact set. The complexity of the resulting optimization program is linear in N. It therefore scales and leads to substantial improvements in computational time over alternative generic algorithms for mixed integer programs, even for relatively small N.

The key idea is that holding the commission fixed, the individual profitability of any agent is independent of the composition. The principal's problem of whether to keep an agent or not at a given commission therefore depends only on whether the agent is profitable at that commission, and not on the composition itself. We can then solve the problem with an inner-outer loop procedure. In the inner loop, we find the optimal composition for any commission by retaining only agents that are profitable at that commission. In the outer loop, we search for the optimal commission over the unit interval. We now formalize the idea.

Equation (10) gives the optimal common commission β for any considered composition profile.

Given β one can invert out the firm optimal salaries from the IR constraint agent-by-agent,

$$\boldsymbol{\alpha}(\beta) : \mathcal{CE}(\boldsymbol{\alpha}, \beta) = \boldsymbol{U}^0 \tag{13}$$

where we note that the optimal salaries $\alpha(\beta)$ are continuous in β. The optimal contracts are now summarized by β. Define agent i's profit contribution at contract β,

$$\pi_i(\beta) = \mathbb{E}[S_i(\beta) - (\alpha_i(\beta) - \beta S_i)] \text{ for all } i \in \mathcal{M}$$

We then show that at the optimum, the profit contribution of any included agent is positive.

Proposition 1 *At the optimal composition-compensations* $(\mathcal{M}^*, \beta(\mathcal{M}^*))$

1. $\pi_i(\beta(\mathcal{M}^*)) \geq 0$ *for all* $i \in \mathcal{M}^*$
2. $\pi_i(\beta(\mathcal{M}^*)) < 0$ *for all* $i \in (1, .., N) \backslash \mathcal{M}^*$

Proof. For the first part, suppose not, and that $\pi_{i'}(\beta(\mathcal{M}^*)) < 0$ for some agent $i' \in \mathcal{M}^*$. Then the firm could fire agent i' and hold the compensation $\mathcal{W}(\beta(\mathcal{M}^*))$ fixed for the remaining agents in $\mathcal{M}^* \backslash i'$. Since none of the remaining agents face different incentives, their profit contributions stay constant and net profits would improve. But that contradicts the optimality of the composition-compensation $(\mathcal{M}^*, \beta(\mathcal{M}^*))$. It follows that $\pi_i(\beta(\mathcal{M}^*)) \geq 0$ for all $i \in \mathcal{M}^*$. Part 2 follows by analogous logic. ■

The intuition is straight-forward: if an agent excluded from the optimal composition can be profitably employed without changing the incentives for the other agents, then that agent can clearly be profitably employed at a commission optimized for the composition including him. If so, the excluded agent belongs to the optimal composition, and we have a contradiction. We can then construct the *conditional value function* that concentrates out the composition problem conditional on β as,

$$\pi(\beta) = \sum_{i \in (1,..N)} \mathbf{1}_{\pi_i \geq 0} \pi_i(\beta). \tag{14}$$

The concentrated value function sums the profit contributions of the agents that are retained in the optimal composition conditional on β. Since the composition problem conditional on β involves integer optimization, one may suspect the conditional value function is discontinuous, precluding a smooth search over β. We now prove the following useful property of the conditional value function:

Proposition 2 *The conditional value function $\pi(\beta)$ is continuous in β.*

Proof. Both the expectation of the sales process in Equation (4) and the concentrated salaries in Equation (13) are continuous functions of β, so it follows that the $\pi_i(\beta)$ is continuous in β for all $i \in (1, .., N)$. The profit contribution of any agent i to $\pi(\beta)$ is the upper envelope of two continuous functions of β, that is $\max\{0, \pi_i(\beta)]\}$, which is itself continuous. Finally, for any β, $\pi(\beta)$ is the sum of continuous functions, and is therefore itself also continuous. ■

4.1 Algorithm

The results above give us the solution algorithm:

1. For a given β, calculate $\pi_i(\beta)$ for all $i \in (1, .., N)$.

2. Search the conditional value function (14) over $[0, 1]$ for the optimal β^*.

At β^*, the optimal composition is $i : \pi_i(\beta^*) \geq 0, i \in (1, .., N)$. The algorithm reduces the joint problem to a search of a continuous function over a compact set, which is a standard optimization problem. Utilizing this characterization enables us to reduce the problem to a unidimensional search over $\beta \in [0, 1]$, rather than a combinatorial search over the the space of 2^N compositions. For N moderately large, this approach can yield a significant improvement in terms of computational time. Note that $\pi(\beta)$ is continuous, piecewise differentiable, and not generally globally concave. We provide some illustrations of its properties for various sizes of N in the appendix.

5 Empirical Application

Our data come from the direct selling arm of the sales-force division of a large contact lens manufacturer in the US (we cannot reveal the name of the manufacturer due to confidentiality reasons).

These data were used in Misra and Nair (2001), and some of the description in this section partially borrows from that paper. Contact lenses are primarily sold via prescriptions to consumers from certified physicians. Importantly, industry observers and casual empiricism suggests that there is little or no seasonality in the underlying demand for the product. The manufacturer employs a direct salesforce in the U.S. to advertise and sell its product to physicians (also referred to as "clients"), who are the source of demand origination. The data consist of records of direct orders made from each doctor's office via a online ordering system, and have the advantage of tracking the timing and origin of sales precisely. Agents are assigned their own, non-overlapping, geographic territories, and are paid according to a nonlinear period-dependent compensation schedule. Pricing issues play an insignificant role for output since the salesperson has no control over the pricing decision and price levels remained fairly stable during the period for which we have data. The compensation schedule for the agents involves salaries, quotas and ceilings. Commissions are earned on any sales exceeding quota and below the ceiling. The salary is paid monthly, and the commission, if any, is paid out at the end of the quarter. The sales on which the output-based compensation is earned are reset every quarter. Additionally, the quota may be updated at end of every quarter depending on the agent's performance ("ratcheting"). Our data includes the history of compensation profiles and payments for every sales-agent, and monthly sales at the client-level for each of these sales-agents for a period of about 3 years (38 months).

The firm in question has over 15,000 SKU-s (Stock Keeping Units) of the product. The product portfolio reflects the large diversity in patient profiles (e.g. age, incidence of astigmatism, nearsightedness, farsightedness etc.), patient needs (e.g. daily, disposable etc.) and contact lens characteristics (e.g. hydrogel, silicone-hydrogel etc.). The product portfolio of the firm features new product introductions and line extensions reflecting the large investments in R&D and testing in the industry. The role of the sales-agent is partly informative, by providing the doctor with updated information about new products available in the product-line, and by suggesting SKU-s that would best match the needs of the patient profiles currently faced by the doctor. The sales-agent also plays a persuasive role by showcasing the quality of the firm's SKU-s relative to that of competi-

tors. While agent's frequency of visiting doctors is monitored by the firm, the extent to which he "promotes" the product once inside the doctor's office cannot be monitored or contracted upon. In addition, while visits can be tracked, whether a face-to-face interaction with a doctor occurs during a visit in within the agent's control (e.g., an unmotivated agent may simply "punch in" with the receptionist, which counts as a visit, but is low on effort).[4]

Misra and Nair (2011) used these data to estimate the underlying parameters of the agent's preferences and environments using a structural dynamic model of forward-looking agents. For our simulations, we use some parameters from that paper, while some are calibrated. We provide a short overview of the model and estimation below, noting differences from their analysis in passing.

5.1 The Model for Sales-Agents

The compensation scheme involves a salary, α_t, paid in month t, as well as a commission on sales, β_t. The sales on which the commission is accrued is reset every R months. The commission β_t is earned when total sales over the sales-cycle, Q_t, exceeds a quota, a_t, and falls below a ceiling b_t. No commissions are earned beyond b_t. Let I_t denote the months since the beginning of the sales-cycle, and let q_t denote the agent's sales in month t. Further, let χ_t be an indicator for whether the agent stays with the firm. $\chi_t = 0$ indicates the agent has left the focal company and is pursuing his outside option. Assume that once the agent leaves the firm, he cannot be hired back (i.e. $\chi_t = 0$ is an absorbing state). The total sales, Q_t, the current quota, a_t, the months since the beginning of the cycle I_t, and his employment status χ_t are the state variables for the agent's problem. We collect these in a vector $\mathbf{s}_t = \{Q_t, a_t, I_t, \chi_t\}$, and collect the observed parameters of his compensation scheme in a vector $\Psi = \{\alpha, \beta\}$. We will use the data in combination with a model of agent behavior to back out the parameters indexing agent's types. The results in this paper are obtained taking these parameters as given.

The index i for agent is suppressed in what follows below. At the beginning of each period, we assume the agent observes his state, and chooses to exert effort e_t. Based on his effort, sales q_t

[4]The firm does not believe that sales-visits are the right measure of effort. Even though sales-calls are observed, the firm specifies compensation based on sales, not calls.

are realized at the end of the period. Sales q_t is assumed to be a stochastic, increasing function of effort, e and a demand shock, ϵ_t, $q_t = q(\varepsilon_t, e)$. The agent's utility is derived from his compensation, which is determined by the incentive scheme. We write the agent's monthly wealth from the firm as, $W_t = W(\mathbf{s}_t, e_t, \varepsilon_t; \mu, \Psi)$ and the cost function as $\frac{de_t^2}{2}$, where d is to be estimated. We assume that agents are risk-averse, and that conditional on $\chi_t = 1$, their per-period utility function is,

$$u_t = u(Q_t, a_t, I_t, \chi_t = 1) = \mathbb{E}[W_t] - r \times \text{var}[W_t] - \frac{de_t^2}{2} \tag{15}$$

Here, r is a parameter indexing the agent's risk aversion, and the expectation and variance of wealth is taken with respect to the demand shocks, ε_t. In the case of a salary + piece-rate of the type considered before, equation (15) collapses to exactly the form denoted in equation (5) for the certainty equivalent. We can thus interpret equation (15) as the nonlinear-contract analogue to the certainty equivalent of the agent under a linear commission. The payoff from leaving the focal firm and pursuing the outside option is normalized to U^0, i.e., $u_t = u(Q_t, a_t, I_t, \chi_t = 0) = U^0$.

In this model, sales are assumed to be generated as a function of the agent's effort, which is chosen by the agent maximizing his present discounted payoffs subject to the transition of the state variables. The first state variable, total sales, is augmented by the realized sales each month, except at the end of the quarter, when the agent begins with a fresh sales schedule, i.e.,

$$Q_{t+1} = \begin{cases} Q_t + q_t & if \quad I_t < R \\ 0 & if \quad I_t = R \end{cases} \tag{16}$$

For the second state variable, quota, we estimate a semi-parametric transition function that relates the updated quota to the current quota and the performance of the agent relative to that quota in the current quarter,

$$a_{t+1} = \begin{cases} a_t & if \quad I_t < R \\ \sum_{k=1}^{K} \theta_k \Gamma(a_t, Q_t + q_t) + v_{t+1} & if \quad I_t = R \end{cases} \tag{17}$$

In above, the new quota is allowed to depend flexibly on a_t and $Q_t + q_t$, via a Korder polynomial basis indexed by parameters, θ_k to capture in a reduced-form way, the manager's policy for updating agent's quotas. The term v_{t+1} is an i.i.d. random variate which is unobserved by the agent in

month t. The distribution of v_{t+1} is denoted $\mathcal{G}_v(.)$, and will be estimated from the data. The transition of the third state variable, months since the beginning of the quarter, is deterministic, augmented by one with the passage of calendar time within the quarter. Finally, the agent's employment status in $(t+1)$, depends on whether he decides to leave the firm in period t. Given the above state-transitions, we can write the agent's problem as choosing effort to maximize the present-discounted value of utility each period, where future utilities are discounted by the factor, ρ. We collect all the parameters describing the agent's preferences and transitions in a vector $\Omega = \{\mu, d, r, \mathcal{G}_\varepsilon(.), \mathcal{G}_v(.), \theta_{k,k=1,..,K}\}$. In month $I_t < R$, the agent's present-discounted utility under the optimal effort policy can be represented by a value function that satisfies the following Bellman equation,

$$V(Q_t, a_t, I_t, \chi_t; \Omega, \Psi) =$$

$$\max_{\chi_{t+1}\in(0,1), e>0} \left\{ \begin{array}{l} u(Q_t, a_t, I_t, \chi_t, e; \Omega, \Psi) \\ +\rho \int_\varepsilon V(Q_{t+1} = Q(Q_t, q(\varepsilon_t, e)), a_{t+1} = a_t, I_t + 1, \chi_{t+1}; \Omega, \Psi) f(\varepsilon_t) d\varepsilon_t \end{array} \right\} \tag{18}$$

Similarly, in month $I_t = R$, the Bellman equation determining effort is,

$$V(Q_t, a_t, R, \chi_t; \Omega, \Psi) =$$

$$\max_{\chi_{t+1}\in(0,1), e>0} \left\{ \begin{array}{l} u(Q_t, a_t, R, \chi_t, e; \Omega, \Psi) \\ +\rho \int_v \int_\varepsilon V(Q_{t+1} = 0, a_{t+1} = a(Q_t, q(\varepsilon_t, e), a_t, v_{t+1}), 1, \chi_{t+1}; \Omega, \Psi) \\ \qquad \times f(\varepsilon_t) \phi(v_{t+1}) d\varepsilon_t dv_{t+1} \end{array} \right\} \tag{19}$$

Conditional on staying with the firm, the optimal effort in period t, $e_t = e(\mathbf{s}_t; \Omega, \Psi)$ maximizes the value function,

$$e(\mathbf{s}_t; \Omega, \Psi) = \arg\max_{e>0} \{V(\mathbf{s}_t; \Omega, \Psi)\} \tag{20}$$

The agent stays with the firm if the value from employment is positive, i.e.,

$$\chi_{t+1} = 1 \text{ if } \max_{e>0} \{V(\mathbf{s}_t; \Omega, \Psi)\} \geqslant 0$$

This completes the specification of the model specifying the agent's behavior under the plan that generated the data. Given this set-up, the structural parameters describing an agent Ω, are estimated in two steps.

Estimation

First, we recognize that once effort, $\widehat{e}_t$ is estimated, we can treat hidden actions as known. The theory implies $\mathbf{s}_t$ is the state vector for the agent's optimal dynamic effort choice. We can use the theory, combined with dynamic programming to solve for the optimal policy function $e^*(\mathbf{s}_t;\Omega)$, given a guess of the parameters Ω. Because $\widehat{e}_t$ is known, we can then use $\widehat{e}_t = e^*(\mathbf{s}_t;\Omega)$ as a second-stage estimating equation to recover Ω. Misra and Nair (2011) implement this approach agent-by-agent to recover Ω for each agent separately. They exploit panel-data available at the client-level for each agent to avoid imposing cross-agent restrictions, thereby obtaining a semi-parametric distribution of the types in the firm.

The question remains how the effort policy, $\widehat{e}_t = \hat{e}(\mathbf{s}_t)$ can be obtained? The intuition used in Misra and Nair is to exploit *the nonlinearity of the contract* combined with panel data for identification. The nonlinearity implies the history of output within a compensation horizon is relevant for the current effort decision, because it affects the shadow cost of working today. Thus, effort is time varying, and dynamically adjusted. The relationship between current output and history is observed in the data. This relationship will pin down hidden effort. Intuitively, the path of output within the compensation cycle is informative of effort. We refer the reader to that paper for further details of estimation and identification.[5] For the counterfactuals in this paper, we need estimates of $\left\{h, k, d, r, U^0, \mathcal{G}_\varepsilon(.)\right\}$. Here, we assume that $\mathcal{G}_\varepsilon(.) \sim N\left(0, \sigma^2\right)$. So, we need $\theta \equiv \left\{h, k, d, r, U^0, \sigma\right\}$. We use the same parameters from Misra and Nair for these, estimating σ from imposing the normality assumption of the recovered demand-side errors from the model. The parameter, k, is not estimated in Misra-Nair. Here, we exploit additional data not used in that paper on the number of calls made by each agent i to a client j in month t, which we denote as, k_{ijt}.

[5] See also, Steenburgh (2008) and Larkin (2010) who note that effort is affected by how far away the agent is from his quota.

We observe k_{ijt} and obtain a rough approximation to k_i as $k_i \approx \frac{1}{T}\sum_t \sum_j k_{ijt}$. The incorporation of k into the model does not change any of the other parameters estimated in Misra-Nair, and only changes their interpretation. We use these parameters for all the simulations reported below.[6]

6 Results

We first discuss the results from the calibration of the agent type parameters. These are reported below. We use a set of 58 agents in our analysis who are all located in one division of the firm's overall salesforce. The numbers we report have been scaled to preserve confidentiality; however, the scaling is applied uniformly and are comparable across agents. For purposes of intuition the reader should consider h and U^0 to be in millions of dollars. So roughly speaking, the median outside option in the data is about $86,400 while the average agent's sales in the absence of effort would be close to a million dollars.

	Parameter					
Statistic	h	k	r	d	σ	U^0
Mean	0.9618	1.0591	0.0466	0.0436	0.4081	0.0811
Median	0.9962	1.0802	0.0314	0.0489	0.3114	0.0864
Min	0.5763	0.2642	0.0014	0.0049	0.0624	0.0710
Max	1.4510	1.8110	0.3328	0.1011	1.5860	0.1032

The plan for the rest of the paper is as follows. We condition on the parameters above and solve for the optimal composition and compensation for the firm using the algorithm described previously. We then discuss these below, simulating two different scenarios. First, we simulate the fully heterogeneous plan where each agent receives a compensation plan (salary + commission) tailored specifically for him or her. We also simulate the partially homogenous contract where the commission rate is common across agents but the salaries may vary across individuals. In all the results presented below, we assume that when an agent is excluded from the salesforce, the territory

[6]Note that given that the data are from only one firm and there is no hire/fire variation, only an upper bound on each agent's outside option is identified. We use these upper bounds as the estimate for U_i^0 in our results reported below. Point estimation of each agent's outside option will require data on workers leaving the firm and their pay elsewhere, and specifying a fuller model of labor market sorting. In simulations we conducted, modest uncertainty the firm has around U_i^0 did not significantly alter the results reported.

provides revenues equal to τh with $\tau = 0.95$, and h is the intercept in the output equation. This assumption reflects the fact that even if a territory is vacated, sales would still accrue on account of the brand or because the firm might use some other (less efficient) selling approach like advertising. We also explored alternative assumptions (e.g. $\tau = 0$ and $\tau = 1$); these results are available from the authors upon request. Qualitatively, the results obtained were similar to those presented below. Below we organize our discussion by presenting details of the optimal composition chosen by the firm under these plans, and then present details of effort, sales and profits.

6.1 Composition

We start with the fully heterogeneous plan as a benchmark. We find all agents have positive profit contributions when plans can be fully tailored to their types. Consequently, the optimal configuration under the fully heterogeneous plan is to retain all agents (the "status quo"). This is not surprising as noted in our 3-agent simulation previously.

Simulating the partially heterogeneous compensation plans, we find the optimal composition in this salesforce would involve letting go of seven salespeople. It is interesting to investigate the characteristics of the agents who are dropped and to relate it to that of the agent pool as a whole. In Figure (1) we plot the joint distribution of the primitive agent types $\left\{h, k, d, r, U^0, \sigma\right\}$ for all agents at the firm. The marginal densities of each parameter across agents is presented across the diagonal. Each point in the various two-way plots along the off-diagonals is an agent, and each two-way shows a scatter-plot of a particular pair of agents types, across the agent pool. For instance, plot [4,1] in Figure (1) shows a scatter-plot of risk aversion (r) versus the cost of effort (d) across all agents in the pool. Plot [1,4] is symmetric and shows a scatter-plot of cost of effort (d) versus risk aversion (r). The seven agents who are not included in the optimal composition are represented by non-solid symbols, highlighted in red. For instance, we see that one of the dropped agents, represented as an "o", has a high risk aversion (1^{st} row), an average level of sales-territory variance (2^{nd} row), an average level of productivity (3^{rd} row), a low cost of effort (4^{th} row), a low outside option (5^{th} row), and a lower than average base level of sales (last row). This agent has a low cost of effort. However, his high risk aversion, his lower outside option, as well as his fit relative

to the distribution of these characteristics across the rest of the agents, implies he is not included in the preferred composition. Figure (1) illustrates the importance of multidimensional heterogeneity in the composition-compensation tradeoff facing the principal, and emphasizes the importance of allowing for rich heterogeneity in empirical incentive settings.

In Figure 2, we plot the location of these salespeople on the empirical marginal densities of the profitability and sales across sales-agents. What is clear from Figure 2 is that there is no a priori predictable pattern in the location of these agents. In some cases, the agents lie at the tail end of the densities, though this does not hold generally. Further, the dropped agents are not uniformly at the bottom of the heap in terms of expected sales or profit contribution under the fully heterogeneous plan. For example, #33, one of the dropped agents, has expected sales of \$1.70MM under the fully heterogeneous plan which would place him/her in the top decile of agents in terms of sales. In addition, he/she is also in the top decile across agents in terms of profitability. However, in his/her case the variance of sales is the highest in the firm, and this creates a large distortion in the contract via the effect it induced on the optimal commission rate (β). Not including this agent allows the firm to improve the contract terms of other agents thereby increasing profits. Other agents are similarly not included on account of some other externality that impacted the compensation contract.

Figure 1 and 2 accentuates the difficulties of ranking agents as desirable or not on the basis of a single type-based metric and the need for a theory of value to assess sales-agents.

6.2 Compensation

We now discuss the optimal compensation implied for the firm under the optimal composition. We compare the fully heterogeneous plan to the partially homogenous plans with and without optimizing composition. Figure (3) plots the density of optimal commission rates under the fully heterogeneous plan along with those for the partially homogenous plans. The solid vertical lines are drawn at the common commission rate for the homogenous plans, with the blue vertical line corresponding to optimizing composition and the black corresponding to not optimizing composition. Looking at Figure (3), we see that the commission rates vary significantly across the sales-agents under the fully heterogeneous plans, going as high as 2.5% for some agents (median commission of about 1.2%).

Figure 1: Joint Distribution of Characteristics of Agents who are Retained and Dropped from Firm under Partially Homogenous Plans

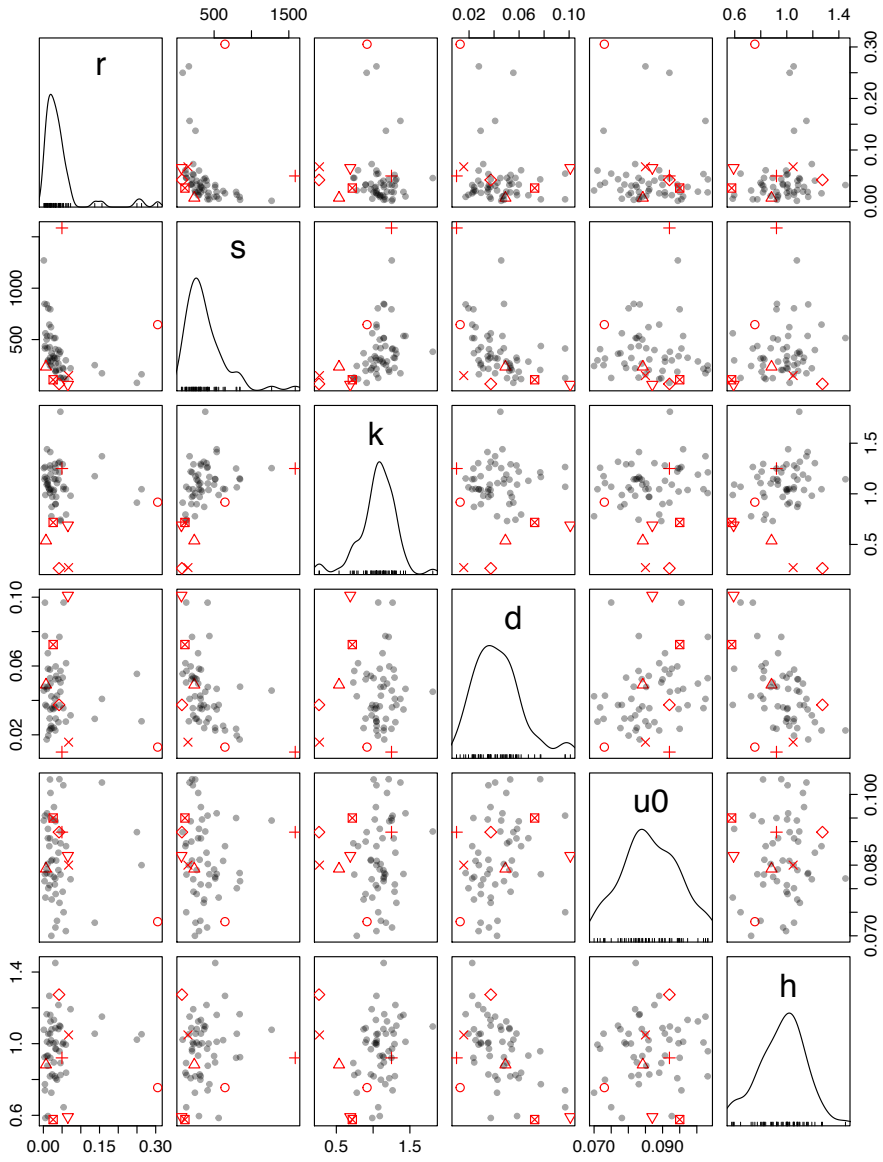
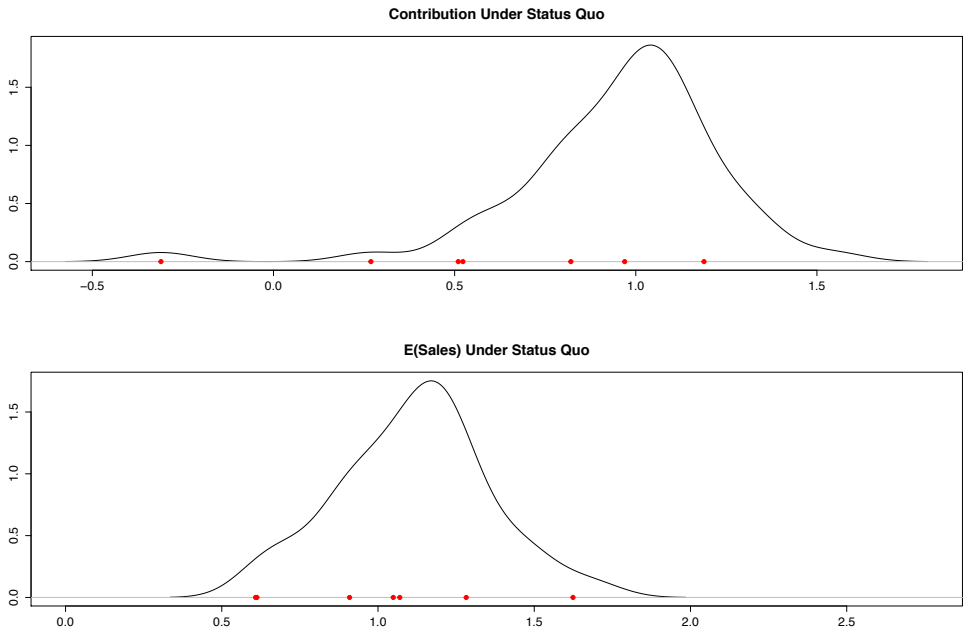


Figure 2: Profitability and Sales of Eliminated Sales Agents

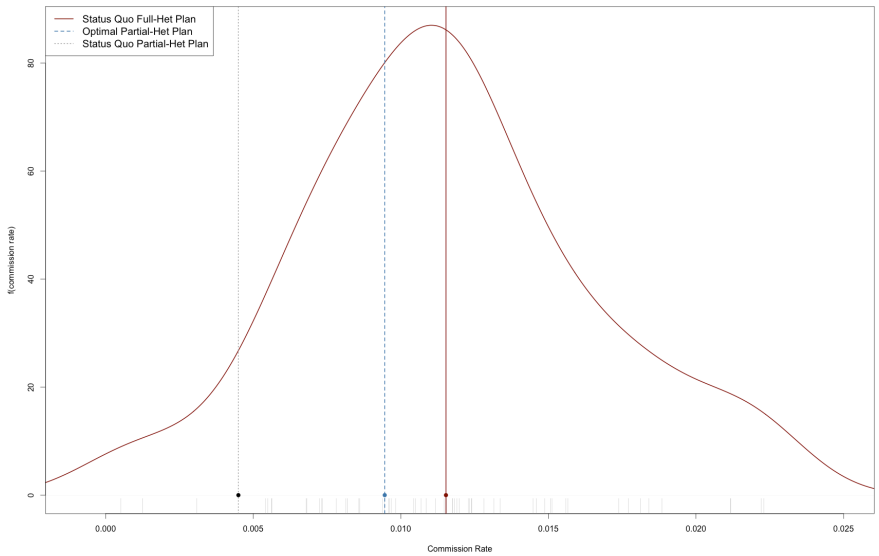


Under the partially homogenous plans, the optimal commission rates are lower. Interestingly, the ability to fine tune composition has significant bite in this setting. In particular, when constrained to not fine tune the salesforce, the firm sets an optimal common commission of about 0.5%. When it can also fine tune the salesforce, the firm optimally sets a higher commission rate of about 0.9%. When the firm is constrained by the compensation structure, the extreme agents (eliminated in the optimal composition) exert an externality that brings the overall commission rate down. By eliminating the "bad" agents, the firm is able to increase incentives. To what extent does does this improve effort, sales and profitability? We discuss this next.

6.3 Effort and Outcomes

The profits for the firm under the fully heterogeneous plan are estimated to be around $60.56MM. We decompose profits with and without homogenous plans, with and without optimizing composition. As noted above, in our data all agents have positive profit contributions and consequently, the optimal configuration is identical to the status quo for compensation plans that are fully heterogeneous. Consequently profits for the fully heterogeneous plan under the optimal configuration

Figure 3: Optimal Commission Rates Under Fully Heterogeneous, Fully Homogenous and Partially Homogenous Plans



	Composition →	
Compensation ↓	*Status Quo*	*Optimal*
Fully Heterogeneous	\$60.56MM	\$60.56MM
Partially Homogenous	\$55.78MM	\$59.18MM

Table 1: Profits under Fully and Partially Heterogenous Plans

and the status quo are identical. This is depicted in the first row of Table (1).

In contrast to the fully heterogeneous compensation structure, there is a significant difference in profit levels when compensation plans cannot be customized. Looking at the above table, partially homogenous plans with the ability to fine-tune composition come very close to the fully-heterogeneous plan in terms of profitability (\$59.18MM compared to \$60.56MM). But partially homogenous plans without the ability to fine-tune the salesforce causes a distortion in incentives, and result in a profit shortfall of \$3.4MM, bringing the total profits down to \$55.78MM.

To decompose the source of profitability differences across the different scenarios, in Figures 4a and 4b we depict the empirical CDF of effort and sales under the three scenarios. The "status-quo" plan is the one that keeps the same composition as currently, but changes compensation. Both the sales and effort distributions under the fully heterogeneous plan fall to the right of the

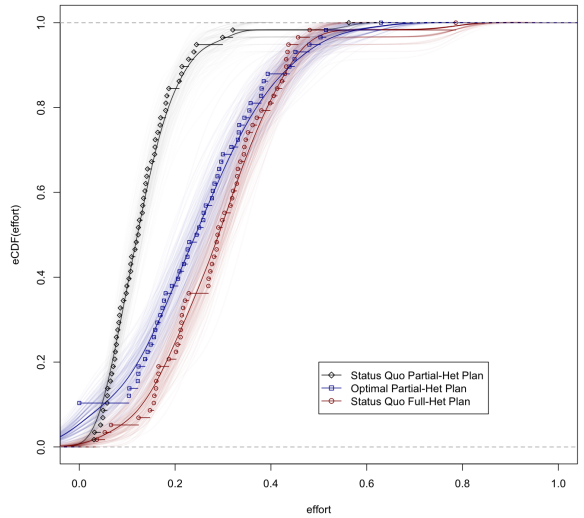
partially heterogeneous plans. However, Kolmogorv-Smirnoff tests show that the distribution of sales and effort under the composition and compensation optimized scenario is *not* statistically different from that under the fully heterogeneous plan. This is striking, since it suggests that by simply altering composition in conjunction with compensation a firm can reap large dividends in motivating effort, even under the constraints of partial homogeneity in contractual terms. This is also why the overall profits under the optimal composition with common commissions is so close to that under heterogeneous plans.

We now assess the extent to which profits at the *individual sales-agent* level under the partially homogenous plan combined with the ability to choose the composition of agents, approximates the profitability under the fully heterogeneous plan (the baseline or best-case scenario). In Figure 6, we plot the profitability (revenues − payout) of each agent under the fully heterogeneous plan on the x-axis, and the profitability under the partially homogenous plan with and without the ability to optimize composition on the y-axis. Solid dots represent profits when optimizing composition, while empty dots represent profits holding composition fixed at the status quo. Each point represents an agent. Numbers are in \$MM-s. Looking at Figure 6, we see the ability to choose composition is important. In particular, the profitability at the agent-level when constrained to partially homogenous contracts and not optimizing composition lies much below the profitability under a situation where contracts can be fully tailored to each agent's type. But, the ability to choose agents seems to be able to mitigate the loss in incentives implied by the constraint to homogeneity. The profitability under the composition-optimized, partially homogenous contracts come very close to that under fully tailored contracts.

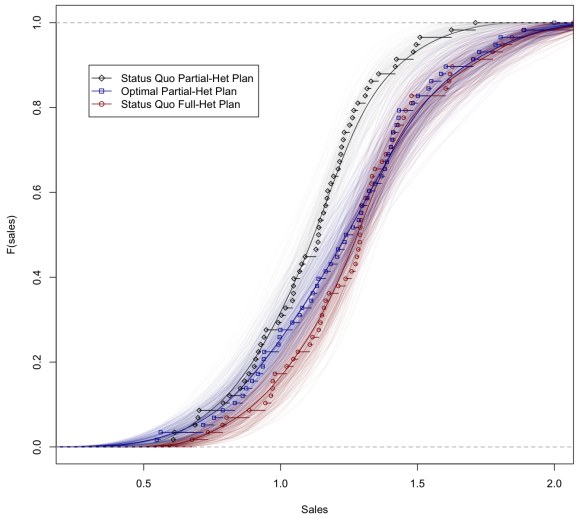
We think this is an important take away. In the real-world, firms can choose both agents and incentives, and not incentives alone. Firms do face constraints when setting incentives. But, our results suggest that the profit losses associated with these constraints are lower when firms are also able to choose the type-space of the agents concomitant with incentives.

Figure 4: Empirical CDF of Implied Effort and Sales Under Different Counterfactual Compensation and Composition Profiles

(a) Empirical CDF of Effort Profiles



(b) Empirical CDF of Expected Sales



Mechanism

The question remains what is the mechanism that enables the firm to come close to the fully heterogeneous plan when it optimizes the composition of its agents? The intuition is straightforward. When constrained to set a homogenous plan, a firm can do much better if the agents it has to incentivize are more homogenous. Consider an extreme case where the firm could find as many agents of any type for filling its positions (no search costs for labor). Then, the firm would first pick the agent from who it could obtain the highest profit (output − payout) under the fully tailored heterogenous contract. It would then fill the N available positions with N replications of that agent. Then, the uniform commission it charges for the salesforce as a whole will be optimal for every agent in the firm. Thus, heterogeneity involves costs. In this sense, an increase in heterogeneity has two roles for a firm constrained to uniform contracts. On the one hand, it increases the chance that high quality agents are in the firm (a positive). But on the other hand, it also increases the level of contractual externalities (a negative). The optimal composition has to balance these competing forces.

To see this more formally, consider a firm that has demand for two agents, which it can fill with A or B type agents. Let Θ index type and suppose type$A$-s generate more profit than B-s when employed at their individually optimal contracts:

$$\Theta_A \neq \Theta_B \quad \text{and} \quad \mathbb{E}\left[\pi\left(\{A\}\right)\right] > \mathbb{E}\left[\pi\left(\{B\}\right)\right]$$

Then, all things held equal, the firm would prefer composition $\{A, A\}$ over $\{B, B\}$,

$$\mathbb{E}\left[\pi\left(\{A, A\}\right)\right] > \mathbb{E}\left[\pi\left(\{B, B\}\right)\right]$$

Now suppose that types are such that though $\Theta_A \neq \Theta_B$, A-s and B-s generate the same expected profit when employed at their individually optimal contracts,

$$\Theta_A \neq \Theta_B \quad \text{but} \quad \mathbb{E}\left[\pi\left(\{A\}\right)\right] = \mathbb{E}\left[\pi\left(\{B\}\right)\right]$$

Then, even though individual profits are the same, the firm would prefer to have the composition $\{A, A\}$ or $\{B, B\}$ over $\{A, B\}$, because composition $\{A, B\}$ generates contractual externalities,

$$\mathbb{E}\left[\pi\left(\{A, A\}\right)\right] = \mathbb{E}\left[\pi\left(\{B, B\}\right)\right] > \mathbb{E}\left[\pi\left(\{A, B\}\right)\right]$$

It is in this sense that the firm has a preference for heterogeneity reduction. Optimal contracting requires the principal to satisfy both incentive rationality and incentive compatibility constraints for its chosen agents. Allowing for agent-specific salaries allows the firm to satisfy incentive rationality for the agents it wants to retain. But the constraint to a common commission implies that incentive compatibility becomes harder to satisfy when the agent pool becomes more heterogeneous. Hence, a firm that can also choose the pool prefers one that is relatively more homogenous, *ceteris paribus*.

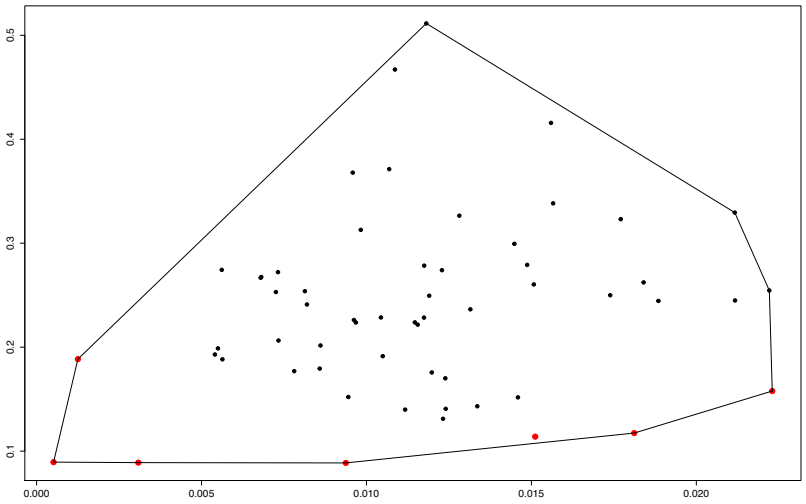
To empirically assess this intuition, we compute two measures of the spread in the type distribution of the salesforce under the optimal partially homogenous contacts with and without the ability to choose agents. Assessing the dispersion in types is complicated by the fact that the type-space is multidimensional. We can separately compute the variance-covariance matrix of types in the salesforce under the two scenarios. To compute a single metric that summarizes the distribution of types, we define a measure of spread, $d_{\mathcal{M}}$, as the trace of the variance covariance matrix of agent characteristics,[7]

$$d_{\mathcal{M}} = \mathrm{tr}\left(\Sigma_{\mathcal{M}}\right) \tag{21}$$

We find that $d_{\mathcal{M}_{\text{StatusQuo}}} = 78,891.6$, and $d_{\mathcal{M}_{\text{Optimal}}} = 51,921.8$, where $d_{\mathcal{M}_{\text{StatusQuo}}}$ is the trace under the optimally chosen partially homogenous plan while retaining all agents in the firm, and $d_{\mathcal{M}_{\text{Optimal}}}$ is the trace under an optimally chosen partially homogenous plan while jointly optimizing the set of agents retained in the firm. We see that the optimal configuration involves about 34.2% reduction in heterogeneity. As another metric, we use $d_{\mathcal{M}} = \det\left(\Sigma_{\mathcal{M}}\right)$. The determinant can be interpreted as measuring the volume of the parallelepiped spanned by the vectors of agent types. To the extent the volume is lower, the spread in types may roughly be interpreted as lesser. We find that the determinant based measure of spread shows a 80.8% decline when the firm can pick

[7]The trace of a matrix is the sum of its diagonals.

Figure 5: Optimal Composition displays a degree of Outlier Aversion



Note: x-axis: β_i, and y-axis: α_i, the commissions and salaries under the fully heterogeneous contracts case. Green dots denote agents retained in the optimal compositions; red dots denote agents dropped in the optimal composition.

its agents and incentives, relative to picking only incentives. Both illustrate that under the optimal strategy, the firm chooses agents such that the residual pool is more homogeneous. While this is intuitive, what is surprising is that its profits under this restricted situation come so close to what it would make under fully heterogeneous plans. This can only be assessed empirically.

To assess the intuition visually, we plot in Figure (5), the salaries and commissions of the entire set of agents when each could be offered his own tailored contract (i.e, the salary + commissions from the fully heterogeneous case). The green dots in Figure (5) denotes the agents who are retained in the optimal composition while the red dots denote the agents who are dropped. Also plotted is the convex hull of the salary/commission points. We see that the optimal configuration exhibits a degree of outlier aversion: the agents dropped from the optimal composition are all on the extremes of the distribution. Note at the same time, that being an outlier does not automatically imply an agent is dropped: we see that some agents on the edges are still retained, presumably on account of their higher abilities or better fit with the rest of the agents.

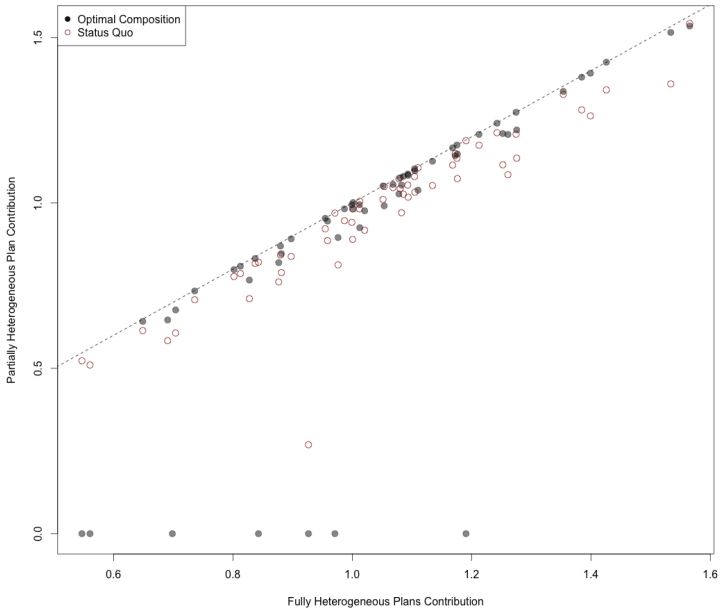
In an important paper, Raju and Srinivasan (1996) make an analogous point, that allowing for heterogeneous quotas in a common commissions setting can closely approximate the optimal

salary + commission based incentive scheme for a heterogeneous salesforce when those quotas can themselves reflect agent specific differences. Our point is analogous, that a firm constrained to a homogenous slope on its incentive contract can come very close to the optimum by picking the region of agent-types that it wants to retain. However, the mechanism we suggest is different. Raju and Srinivasan (1996) suggest addressing the problem of providing incentives to a heterogeneous salesforce by allowing for additional heterogeneity in contract terms. We suggest addressing the problem of setting incentives to a heterogeneous pool of agents by making the salesforce more homogenous. In another contribution, Lal and Staelin (1986) and Rao (1990) show that a firm facing a heterogeneous salesforce can tailor incentives to the distribution of types it faces by offering a menu of salesforce plans. Their approach uses agents' self-selection into plans as the mechanism for managing heterogeneity, and requires the firm to offer a menu of contracts taking the salesforce composition as given. In our model, the firm offers only one contract to an agent, but chooses which agent to make attractive contracts to (thus, the margin of choice for agents in our model is not over contracts but over whether to stay in the firm or leave). Our model endogenizes the salesforce's composition and may be seen as applying to contexts where offering a menu of plans to employees to choose from is not feasible, or desired. We think the three perspectives outlined above for the practical management of heterogeneity in real-world settings are complementary to each other. In the section below, we discuss the latter mechanism further.

6.4 Sorting Agents into Divisions

We now discuss whether we can further improve the management of heterogeneity in the firm by sorting agents into divisions. We consider a salesforce architecture in which the firm creates $|\mathcal{J}|$ divisions, and assigns each agent it retains to one of the $|\mathcal{J}|$ divisions. The divisions correspond to different compensation profiles. We allow each division to have its own commission, but require that all agents within a division are given the same commission. Salaries are allowed to be heterogeneous as before. Such architectures are commonly observed in the real-world. For instance, salesagents targeting large "key accounts" may be assigned into a division which offers more incentive pay, while those targeting smaller clients may be in a division that offers more salary than commission. Or

Figure 6: Profitability at the Individual Sales-agent Level under Fully Heterogeneous Plan and Partially Homogenous Plan with and without Optimized Composition.



alternatively, salesagents targeting urban versus rural clients may be in two different divisions each with its own commission scheme. But the observed empirical fact is that commissions are invariably the same within a division. This architecture reflects that.

For each value of $|\mathcal{J}|$ we solve simultaneously for the match between agents and divisions and the optimal commissions across divisions, along with the optimal salaries across agents given their assignment to a division. Formally, we solve the following modified bi-level optimization problem,

$$\max_{\mathcal{M}_j} \Pi = \sum_{j \in \mathcal{J}} \int \sum_{i \in \mathcal{M}_j} \left(S_i - \mathcal{W}_{\mathcal{M}_j}(S_i) \right) d\mathcal{F}(S_i|e_i), \quad st.,$$

$$\mathcal{W}_{\mathcal{M}_j} = \arg \max_{\mathcal{W} \in \mathbb{W}_N} \int \sum_{i \in \mathcal{M}_j} \left(S_i - \mathcal{W}(S_i) \right) d\mathcal{F}(S_i|e_i), \quad \text{(IR,IC)}$$

$$\bigcup_{j \in \mathcal{J}} \mathcal{M}_j = \mathbb{M}_N$$

where the last "adding-up" constraint ensures that a given agent is either assigned to one of $|\mathcal{J}|$ contracts. The incentive compatibility and rationality constraints IR and IC are not written out

explicitly for brevity. In the final solution to above, the set $\mathcal{M}_j$ assigns to each agent i, a number $\{0, 1, .., j, .., |\mathcal{J}|\}$, where 0 implies the agent is dropped from the firm, and $j > 0$ implies the agent is assigned to division j with wage contract $\mathcal{W}_{\mathcal{M}_j}$. Our goal is to assess empirically how many divisions ($|\mathcal{J}|$) are required to fully span the heterogeneity and to come close to the profits under the fully heterogeneous case. Additionally, we want to assess the extent to which the ability to choose agents interacts with this mechanism for managing heterogeneity.

In Figure (7) we report on the results in which we simulated the profits to the firm from creating upto $|\mathcal{J}| = 6$ divisions. The x-axis of Figure (7) plots the number of divisions considered ($|\mathcal{J}|$). The y-axis of Figure (7) plots the total profits to the firm for each $|\mathcal{J}|$. Each point corresponds to solving the modified bi-level problem above for the corresponding value of $|\mathcal{J}|$. The green-line shows the profit profile in which we allow the firm to sort agents into divisions, but do not allow the firm to optimize the composition (i.e., in the bi-level program above, we do not allow $j = 0$ as an option). The blue-line in the figure shows the profit profile in which the allow the firm to sort agents into divisions and allow the firm to optimize the composition as described above. The difference in the profits under the blue versus the green lines indicates the extent to which composition choice adds to profitability over and above the ability to sort agents into divisions.

We first discuss the situation where we allow the firm to sort agents into divisions but do not allow the firm to choose composition. The top horizontal line in Figure (7) represents a profit of \$60.56M, the maximum profit possible under the fully heterogeneous contract (see Table (1)). Looking at the green-line in Figure (7), we see that even without the ability to choose composition, the firm is able to come very close to this value with as less as 6 divisions. Even two divisions do a remarkably good job of managing heterogeneous incentives − profits under the green-line for the 2-division case are more than \$59M. Thus, one empirical take-away is that a small amount of variation in contract terms seems to be sufficient to manage a large amount of heterogeneity in the firm, at least in the context of these estimates.

We now discuss the situation where we allow the firm to sort agents into divisions and to optimize its composition. We see the results are similar to the previous case, but the firm is able to achieve

a higher level of profit gains with fewer divisions (the blue-line is always above the green-line). Thus, the ability to choose composition has bite even when one allows for sorting into divisions. With $|\mathcal{J}| = 6$ divisions, we find the firm ends up dropping 5 agents from the optimal composition (compared to 6 agents with only one division). Thus, allowing for sorting does not automatically imply that composition choice is not needed − the right perspective is that sorting and composition-choice are two strategies to manage heterogeneity, and when used in combination, unlock powerful complementaries in the provision of firm-wide incentives.

Finally, we show that heterogeneity reduction plays a role in improving profitability with sorting. In Figure (8) we plot the log-distortion implied by the divisions against the number of contracts offered. The distortion metric is simply a way to summarize the average heterogeneity within a division. It captures the mean squared deviation from the average salesperson across divisions. Formally, let θ_{ij} denote the 6×1 vector of characteristics for agent i allocated to division $j > 0$ and let $\theta_{ij}^{(r)}, r = 1, .., 6$, denote the r^{th} element of θ_{ij}. Denoting N_j as the number of agents allocated to division j, let $\bar{\theta}_j^{(r)} = \sum_{i=1}^{N_j} \theta_{ij}^{(r)}/N_j$ be the average value of characteristic r inside division j. We define the distortion as,

$$d_{\mathcal{M}} = \min_{r} \frac{1}{|\mathcal{J}|} \sum_{j=1}^{|\mathcal{J}|} \left(\theta_{ij}^{(r)} - \bar{\theta}_j^{(r)}\right)^2 \tag{22}$$

In Figure (8), the lower line (red) corresponds to the model which allows for composition to be optimized jointly with division-specific commissions, while the top line (blue) ignores the composition aspect. As one would expect, as more contracts are added to the compensation structure, the distortion falls but allowing the firm to manage composition results in a more significant reduction. In essence, with a small number of contracts, the firm finds it optimal to eliminate the outlying agents and use the increased flexibility to better compensate those that are retained. As a result the heterogeneity in the retained agent pool is managed much better. Ultimately, as the number of contracts increase to match the total number of agents, the two curves would coincide. That is, if every agent got a customized contract there would be no distortion. The broad point is that sorting and composition choice together enable very effective management of heterogeneity even with restricted contracts.

Figure 7: Performance of Divisions in Managing Heterogeneity

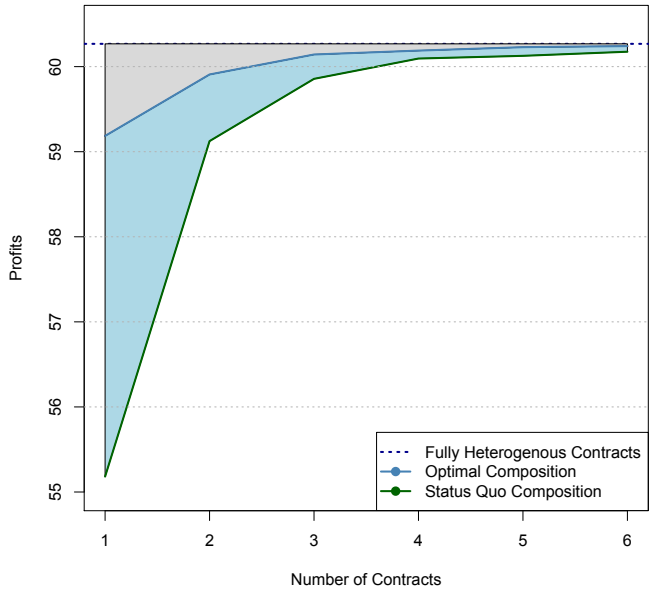
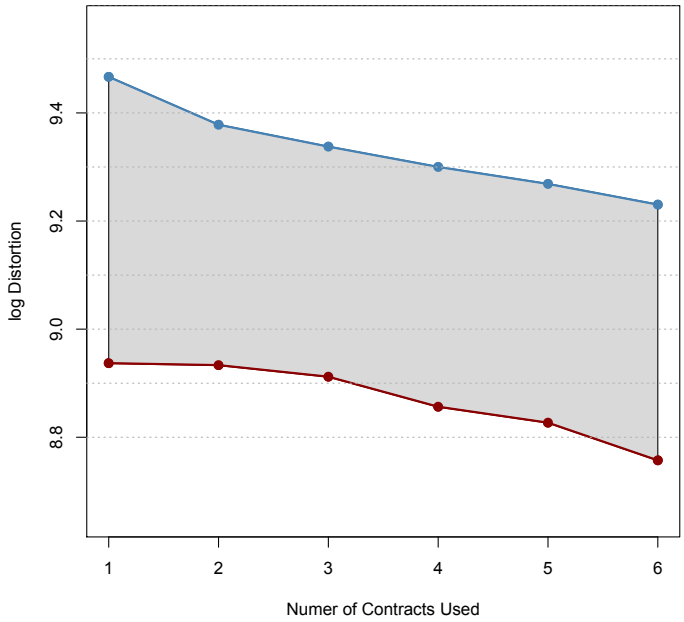


Figure 8: Heterogeneity Reduction in the Saleforce with Many Divisions



Discussion

We developed the above exercise under the assumption that the firm knows each agent's type perfectly and can sort agents into divisions based on that knowledge. As mentioned above, the theory has emphasized an alternative mechanism in which the firm offers a menu of contracts to all and the agent self-selects into one of the offered contracts based on his unknown type (this is analogous to nonlinear pricing). This strategy helps manage heterogeneity when the firm does not know types perfectly and emphasizes adverse-selection as the main difficulty in contract design, as opposed to the moral hazard we emphasize. In practice, it is likely that both are at play in many real-world contexts. Optimal contract design with both adverse-selection and moral hazard is beyond the scope of this paper. In salesforce contexts, we believe the approach we have outlined above is more realistic than self-selection contracts. First, unlike nonlinear pricing, self-selection contracts are rarely observed in salesforce compensation (perhaps due to concerns with dynamic signaling − if an agent chooses a contract with low commissions, he signals his type to the principal which can be used to update his contracts in subsequent periods). The more common observation is of salesforce divisions and of assignment of agents into divisions. Second, adverse-selection in salesforce settings is usually addressed by monitoring, probation and training. New hires are often placed on a salary-only probation period in which their performance is observed. The employment offer is made full-time conditional on satisfactory performance in the probation period. New hires are also provided significant sales training during the probation period and asked to "shadow" an established sales-rep where real-time training is imparted and performance on the field is observed. This monitoring helps the firm assess agent types before full-time offers are made. Thus, in our view, for long-run salesforce composition and compensation with full-time salesagents, adverse selection may be a second-order consideration. A limitation of our model is that it does not apply to the interesting dynamics outlined above associated with new employee hiring and learning.

Finally, if the firm does not know types perfectly, the profits it can make when offering a menu of divisions is strictly lower than the profits it can make when it knows types perfectly and can assign each type to its preferred division (as in the simulations above). We reported above that

when types are known perfectly, firms still gain from the ability to choose composition. We interpret this as implying that even if a menu of contracts are offered, the ability to choose agents that we emphasize, will still have bite in terms of profits in the context of our empirical example.

6.5 Parameter Uncertainty and Robustness

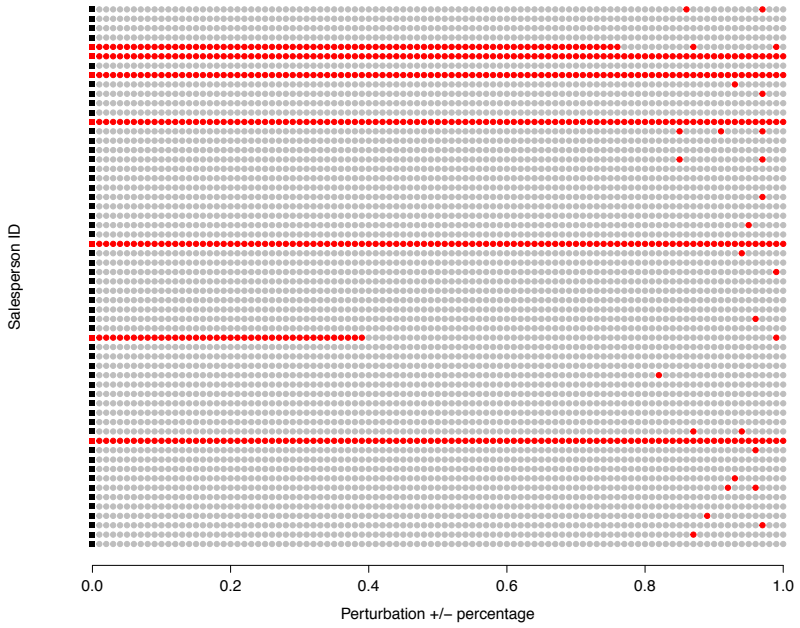
A primitive assumption in our analysis is that the firm knows the agents' types with complete certainty. While this is a standard assumption in principal-agent models with pure moral hazard, it poses some relevant questions for our analysis: Do the results continue to hold if the firm only has access to "estimates" of the agents' types? If the results are different, what are the nature and magnitude of the differences? To answer these, we implemented extensive Monte Carlo simulations in which the firm recognizes it has access to parameters of agent types that are estimated with error, and maximizes an expected profit function that integrates out this estimation error.[8]

Under the assumption that firms acknowledge the presence of estimation error, we need a way of incorporating the parameter uncertainty into the firm's decision making process. We assume the firm has access to the sampling distribution of the estimates. The firm then uses this information to integrate out the uncertainty and maximizes expected profits. We simulate various levels of uncertainty by perturbing the parameters we estimated. We add to each point estimate a noise term that is normally distributed with mean zero and with standard deviations that vary so that the effective range the firm believes each parameter can lie in goes from $\pm 1\%$ to as much as $\pm 100\%$ of the point estimate. This corresponds to situations where is firm is able to estimate the parameter with a high level of precision, to situations where the parameter estimates are no longer significant at the $\alpha = 0.05$ level (i.e., range contains zero). We used these perturbed parameters to compute the expected profits faced by the firm. We then use the same optimization technology described in the paper to compute the optimal compensation and composition under these simulated scenarios. All comparisons to the status quo and heterogeneous plan outcomes are also based on the appropriate expected profit functions.

[8] For the sake of brevity, we do not include complete details of our simulations here. They are available from the authors upon request.

Our simulations show the inclusion of parameter uncertainty does not alter our results qualitatively. We find the composition results are robust across simulations, in that, the same agents are eliminated from the analysis as before in most. To see this, Figure (9) plots the retention or exclusion of each agent in the optimal composition as a function of the degree of perturbation to the parameter estimate. The agents' ID-s are on the y-axis and the degree of perturbation is on the x-axis (range $\pm 1\%$ to $\pm 100\%$). Each column represents a perturbation level, and each square in a vertical column represents an agent in the configuration. Agents dropped in the optimal configuration are represented in red. The first column plots the original results with no perturbation to the parameters. Each column thereafter plots the optimal composition found by maximizing expected profits (subject to appropriate IR and IC constraints) given a certain level of perturbation. Expected profits are found by Monte Carlo simulation with $R = 1,000$ draws over the parameter range. Looking at Figure (9), we see the number of retained agents is fairly stable (usually around the same level as the original result of 7 eliminations) and the identity of those dropped from the pool is roughly preserved. When the perturbation error is substantial ($> \pm 80\%$) there appear to be frays in the optimal composition. Even so, some agents continue to be (optimally) excluded in the solution set, suggesting that the broader point that composition choice can help mitigate externalities under rigid contracts continues to hold even under extreme parameter uncertainty. In another set of Monte-Carlo simulations (not reported here) when we *resample* the set of agents with replacement, the results vary more significantly, suggesting that varying the heterogeneity in the composition is more relevant to the profitability of the firm, than is the estimation error. Overall, these simulations show our findings above are driven by meaningful differences in agent types (i.e. by heterogeneity) not by parameter uncertainty *per se*. Looking at profits, we find the level of profits between across the sets of results are not very different − even when the parameters are perturbed by upto $\pm 100\%$, the profit difference between those presented in the paper and the counterfactual differ by only around 7%. Finally, the relative profits show the same patterns as before, with the homogeneous plan faring the worst and the optimal composition-based plan coming fairly close to the fully heterogeneous plan.

Figure 9: Identity of Agents Dropped from the Optimal Composition as a Function of the Percentage by which Estimated Parameters are Perturbed



7 Conclusions

We consider a situation where a firm that is constrained to set partially homogenous contracts across its agent pool can optimize both its composition and its compensation policy. We find that the ability to optimize composition partially offsets the loss in incentives from the restriction to uniform contractual terms. Homogeneity also implies a particular type of contractual externality within the company. The presence of an agent in the firm indirectly affects the welfare and outcomes of another through the effect he induces on the common element of contracts. This externality exists even in the absence of complementarity in output across agents, team production, common territories or relative incentive terms. Simulations and an application to a real-world salesforce suggest that the ability to choose composition has empirical bite in terms of payoffs, sales-effort and sales.

The paper explores the consequences of uniformity, but not the reasons for uniformity in contracts within firms. Motivations for uniformity could be sales-agent inequity aversion, concerns for fairness in evaluation, preferences for simplicity, or different kinds of menu costs. In some survey

evidence, Lo et al. (2011) conduct field-interviews with managers at industrial firms in four sectors (namely, electrical and non-electrical machinery, transportation equipment and instruments), and report the two main reasons managers cite for not using agent-specific salesforce compensation plans are (a) computational costs of developing complex plans, and, (b) costs associated with managing ex post conflict amongst salesagents induced by differential evaluation. Relatedly, in a survey of 130 business-format franchisors, Lafontaine (1992) reports that 73% of surveyed franchisors choose uniform royalty rates due to reasons of consistency and fairness towards franchisees, and 27% reported choosing uniformity because it reduces the transaction costs of administering and enforcing contracts. It seems therefore that fairness and menu costs play a large role in driving such contract forms.[9] Notwithstanding the reasons, the fact remains that the ability to choose agents and the restriction to partially homogenous contracts is pervasive in real-world business settings. However, principal-agent theory is surprisingly silent on both endogenizing the composition of agents, and exploring the consequences of uniformity. We hope our first-cut on the topic will inspire richer theory and empirical work on the mechanisms causing firms to choose similar contracts across agents, and on the consequences of these choices.

We abstracted away from hiring and from the principal's policies for learning new hires' types. Accommodating these complicates the model by introducing dynamics, but does not change our main point about contractual externalities and the codependence of compensation and composition when contracts cannot be tailored. In our data, we do not have a way of estimating the distribution of worker types in the population or the distribution of search costs for labor amongst firms in this market, both of which are critical inputs to a credible empirical model of labor market sorting.

[9]We conducted interviews with salespeople and sales managers to understand why salaries are typically heterogeneous, while commissions are invariably homogenous in sales organizations. The common story we have heard is as follows. Salaries are typically indexed against those at a hired employee's previous job (typically set as a percentage raise). Thus, they reflect agents' outside options. Agents perceive the differences as fair because the variation is justified by managers as the costs to "beat" competitive salaries to hire their colleagues. There is some evidence in the psychology literature that agents's perceive variation as fair when it is linked to "justifiable" costs. For example, Kahneman et al. (1986) document that agents do not perceive price discrimination across consumers as unfair if they think it derives from differences in costs as opposed to the desire to extract more surplus from those with higher valuations. On the other hand, commissions (and other incentive based pay) reflect a percentage payout to an agent of revenues brought into the firm. A dollar in revenue brought in by an agent A is equally valuable to the principal as a dollar brought in by another agent B; because of this, it becomes difficult for the principal to justify why A and B are rewarded different proportions of the dollar as commissions. Such a policy is typically seen as "unfair". Clearly, the perception of fairness is linked to the deeper psychology of how human beings evaluate these kinds of tradeoffs.

With access to better data, an extension of this sort could be pursued. The reader should note that such competition in contracts across firms have relatively been understudied in empirical work. Finally, another margin along which the principle may manage heterogeneity is to optimize the match between agents and territories (e.g., Skiera and Albers 1998). Analyzing this matching problem while endogenizing the compensation contract is outside the scope of this paper, but is the subject of our ongoing work.

While our context is salesforce compensation, similar ideas to the one explored here arise in other contexts of interest to Marketing. One area is joint choice of consumers and promotions. For instance, Belloni et al. (2012) discuss an algorithm that enables a University to jointly choose a desirable mix of students and the level of scholarships required to attract them. The complication associated with salesforce compensation relative to these situations is the presence of moral hazard. To the extent that we discuss the implications of endogenizing the mix of agents at a firm, we believe our analysis motivates development of richer empirical models of the joint choice of who and how to offer product options to consumers in Marketing and Economics.

8 References

1. Albers, S. and M. Mantrala. (2008). "Models for Sales Management Decisions," Handbook of Marketing Decision Models.

2. Ackerberg, D. and Botticini, M. (2002). "Endogenous Matching and the Empirical Determinants of Contract Form," Journal of Political Economy, 110(3), 564-591.

3. Armstrong, M. (1996). "Multiproduct Nonlinear Pricing," Econometrica, 64, pg. 51-75.

4. Bandiera, O., Barankay, I., and Rasul, I. (2007). "Incentives for Managers and Inequality Among Workers: Evidence from a Firm-level Experiment," Quarterly Journal of Economics, (May), 729-73.

5. Basu, A., R. Lal, V. Srinivasan and R. Staelin (1985). "Sales-force Compensation Plans: An Agency Theoretic Perspective," Marketing Science, 8 (3): 324-342.

6. Belloni, A., Lovett, M., Boulding, W. and Staelin. (2012). "Optimal Admission and Scholarship Decisions: Choosing Customized Marketing Offers to Attract a Desirable Mix of Customers," Marketing Science, 31 (4), 621-636.

7. Ichniowski, C., Shaw, K. and Prennushi, G. (1997). "The Effects of Human Resource Management Practices on Productivity,"American Economic Review, 86,291-313.

8. Desai, P. S. and Srinivasan, K. (1995). "Demand Signaling under Unobservable Effort in Franchising: Linear and Non-linear Price Contracts," Management Science 41(10), 1608-23.

9. Godes, D. and Mayzlin, D. (2012). "Using the Compensation Scheme to Signal the Ease of a Task," working paper, University of Maryland.

10. Green, J. R., and Stokey, N. L. (1983). "A Comparison of Tournaments and Contracts," Journal of Political Economy 91(3), 349-64.

11. Hamilton, B, J. Nickerson, and H. Owan. (2003). "Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation," Journal of Political Economy 111, no. 3:465–97.

12. Holmstrom, B. (1979). "Moral Hazard and Observability", *Bell Journal of Economics*, Vol. 10.

13. Holmstrom, B. (1982). "Moral Hazard in Teams," Bell Journal of Economics 13, no. 2:324–40.

14. Holmstrom, B. and P. Milgrom. (1987). "Aggregation and Linearity in the Provision of Intertemporal Incentives," Econometrica, 55, 303-328.

15. Joseph, K. and Kalwani, M. (1992). "Do Bonus Payments Help Enhance Sales-force Retention?" Marketing Letters, 3 (4): 331-341.

16. John, G. and Weitz, B. (1989). "Salesforce Compensation: An Empirical Investigation of Factors Related to Use of Salary Versus Incentive Compensation," Journal of Marketing Research, 26, 1-14.

17. Kalra, A. and Shi, M. (2001). "Designing Optimal Sales Contests: A Theoretical Perspective." Marketing Science, 20(2), 170-193.

18. Kahneman, K., Knetsch, J.L, Thaler, R. (1986). "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," American Economic Review, 76(4), pp. 728-741, September.

19. Kandel, E., and E. Lazear. (1992). "Peer Pressure and Partnerships," Journal of Political Economy, 100, no. 4:801–17.

20. Lafontaine, F. (1992). "How and Why Do Franchisors Do What They Do: A Survey Report," in Franchising: Passport for Growth and World of Opportunity (Patrick J. Kaufmann ed.), Sixth Annual Proceedings of the Society of Franchising.

21. Lafontaine, F. and Blair, R. (2009). "The Evolution of Franchising and Franchising Contracts: Evidence from the United States," Entrepreneurial Business Law Journal, Vol. 3.2, pp. 381-434.

22. Lal, R. and R. Staelin. (1986). "Salesforce Compensation Plans in Environments with Asymmetric Information," Marketing Science 5(3), pg. 179-198.

23. Lal, R. and V. Srinivasan. (1993). "Compensation Plans for Single- and Multi-Product Salesforces: An Application of the Holmstrom-Milgrom Model," Management Science, 39 (7), 777-793.

24. Lazear, E. and Rosen, S. (1981). "Rank-Order Tournaments as Optimum Labor Contracts," Journal of Political Economy 89, 841-864.

25. Lazear, E. (2000a). "Performance Pay and Productivity," American Economic Review 90, 5:1346–61.

26. Lazear, E. (2000b). "The Power of Incentives," American Economic Review, *P&P* 90:2: 410-414.

27. Larkin, I. (2010). "The Cost of High-Powered Incentive Systems: Gaming Behavior in Enterprise Software Sales," working paper, Harvard Business School.

28. Lim, N., Ahearne, M. J. and Ham, S. H. (2009). "Designing Sales Contests: Does the Prize Structure Matter?" Journal of Marketing Research, 46, 356-371.

29. Lo, D., Ghosh, M. and Lafontaine, F. (2011). "The Incentive and Selection Roles of Sales Force Compensation Contracts," Journal of Marketing Research, 48(4), pp. 781-798.

30. Mantrala, M., P. Sinha and A. Zoltners. (1994). "Structuring a Multiproduct Sales Quota-Bonus Plan for a Heterogeneous sales-force: A Practical Model-Based Approach" Marketing Science, 13(2), 121-144.

31. Milgrom, P. and Roberts, J. (1990). "The Economics of Modern Manufacturing: Technology, Strategy, and Organization," American Economic Review, 80, 511-528.

32. Misra S., A. Coughlan and C. Narasimhan (2005). "Sales-force Compensation: An Analytical and Empirical Examination of the Agency Theoretic Approach," Quantitative Marketing and Economics, 3(1), 5-39.

33. Misra S., E. Pinker and R. Shumsky (2004). "Salesforce design with experience-based learning," IIE Transactions, 36(10), pp. 941-952

34. Misra S. and H. Nair. (2011) "A Structural Model of Sales-Force Compensation Dynamics: Estimation and Field Implementation," Quantitative Marketing and Economics, 9(3), pp. 211-257

35. Mookherjee, D. (1984). "Optimal Incentive Schemes with Many Agents," Review of Economic Studies 51, 433-446.

36. Pendergast, C. (1999). "The Provision of Incentives within Firms," Journal of Economic Literature, 37(1), 7-63.

37. Raju, J. S., and V. Srinivasan. (1996). "Quota-based Compensation Plans for Multi-territory Heterogeneous Sales-forces," Management Science 42, 1454-1462.

38. Rao, R. (1990). "Compensating Heterogeneous Sales-forces: Some Explicit Solutions," Marketing Science, 9(4), 319-342 41.

39. Rochet, J-C., Chone, H. (1998). "Ironing, Sweeping and Multidimensional Screening," Econometrica, 66, pg. 783-826.

40. Rotemberg, J. and G. Saloner. (2000). "Visionaries, Managers, and Strategic Direction," Rand Journal of Economics, 31, Winter, 693-716.

41. Skiera, B., and Albers, S. (1998). "COSTA: Contribution Maximizing Sales Territory Alignment," Marketing Science, 17, 196-214.

42. Steenburgh, T. (2008). "Effort or Timing: The Effect of Lump-sum Bonuses," Quantitative Marketing and Economics, 6, 235-256.

43. Zoltners, A., P. Sinha and G. Zoltners. (2001). "The Complete Guide to Accelerating Salesforce Performance," American Management Association, New York.

A Behavior of Conditional Value Function with N

We have simulated some data to illustrate some of the properties of $\pi(\beta)$. Agents are generated by spreading lognormal noise around parameters $(d, r, U^0) = (1, 2, 0)$. We then plot $\pi(\beta)$ in Figure (10) for various sizes of N.

In the first quadrant, the conditional value function is seen to be continuous and piecewise differentiable. The kinks are at the points where some agent's IR constraint just binds and the agent either enters or exits the composition. At all other points, the criterion is a sum of continuous and differentiable functions. Though the conditional value function has a clearly accentuated maximum in $\beta \in [0, 1]$, this is not a general feature of the problem. The reader may note that the criterion has two local maxima, the second being slightly above 0.2. Multiple maxima is a general feature of the problem.

The complexity of the algorithm is linear in N since only the profit contributions of each agent must be calculated at each iteration. For all N in this example, the optimization is executed in less than $\frac{1}{100}$th of a second using standard numerical derivatives based methods of optimization. In the south east quadrant, the algorithm allows a directed search of a space of 2^{5000} possible compositions in fractions of a second.

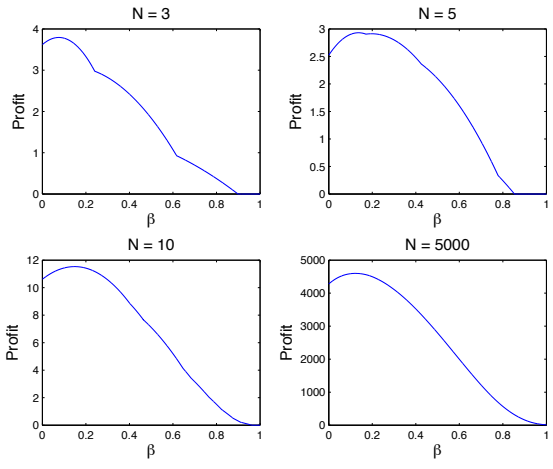


Figure 10: Criterion Function for Various Composition Sizes