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

Sanjog Misra Harikesh S. Nair Øystein Daljord

First version: June 2012. This version: February 2013

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 partially the loss in incentives from the restriction to uniform contracts. We conjecture this may hold more broadly.

* *Misra*: Anderson School of Management, UCLA, `sanjog.misra@anderson.ucla.edu`; *Nair*: Graduate School of Business, Stanford University, `harikesh.nair@stanford.edu`; *Daljord*: Graduate School of Business, Stanford University, `daljord@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 seminar participants at MIT-Sloan, Stanford-GSB, UBC-Saunder, UC-Davis-GSM, UT-Austin-McCoombs, UToronto-Rotman, and at the 2012 IOFest, Marketing Dynamics and QME conferences for very useful comments and discussions. 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, 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.

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

[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."

firm cannot offer high commissions to the less risk averse agents. 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, we need to enlarge the contracting problem 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 on all respects except for their risk aversion, and there are three agents, A, B and C, who could be employed, with C being the most risk averse type. Retaining C in the firm implies the common commission rate the firm would set with C is lower than without, because C needs more insurance than A and B. It could then be that A and B are worse off with C in the firm (lower commissions) than without. 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). To see this, note that in the example above, it may be optimal for the principal to rank 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 well 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 it so happens that C, the most risk averse, is also 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 higher loss in the level of 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 metric.

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 assess how contract form changes when the distribution of ability changes. The answer to this is dependent on the distribution of heterogeneity in the agent pool, and hence is inherently an empirical question. To the best of our knowledge, the interaction between composition and compensation has not been emphasized in the existing principal agent theory and salesforce compensation literature, which typically formulates the principal's problem as choosing the optimal incentive contract taking the set of agents in its pool as given, and not the the endogenous choice by a principal of both agents and incentives simultaneously. The contractual externality we identify also persists when agents have exclusive territories and there are no across-agent complementarity or substitution effects in output. It is thus distinct from across-agent effects induced by relative performance schemes or team-selling, two other contexts which have been emphasized in the literature wherein one agent's characteristics or actions substantively affects another's welfare (Holmstrom 1982; Kandel and Lazear 1992; Hamilton et al. 2003; Misra, Pinker and Shumsky 2004).

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. Following Misra and Nair (2011), we use 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 prespecified 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. Following Misra and Nair (2011), we combine this identification strategy with a structural model of agent's optimization behavior to recover the primitives underpinning agent types. We then 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^M possible salesforce configurations from amongst a pool of M potential agents, and solving for the optimal common incentive contract for the chosen pool (M is around 60 in our application). We find that the recently developed "cross entropy" method (Rubenstein 1997; De Boer et al. 2005) is a practical and reliable way to solve this large-scale optimization problem in realistic settings.[2]

We then use the model to simulate counterfactual contracts and agent pools. We explore to what extent a change in composition of the agents affects the nature of optimal

[2]We experimented with a variety of other optimization methods including simulated annealing and genetic algorithms and found the cross entropy method to be significantly superior in our setting in terms of both speed and reliability.

compensation for those agents, and quantify the profit impact of jointly optimizing over composition and compensation. As a useful by-product of the model, we also simulate a metric which quantifies the value of each potential sales-agent in the pool to the firm. The key insight underlying our metric is that the value of an agent to the firm is not simply the present discounted value of the sales he generates minus the compensation to be paid out to him. Rather, our metric recognizes that the value of an agent has to be defined relative to a counterfactual in which we ask how outcomes would look if the agent were dropped from the firm. A complete counterfactual would take into account that dropping an agent from the firm will change the distribution of sales and payouts for *every* other agent in the firm through its effect on the changed commission the firm would charge subsequently to its residual agent pool. We simulate this counterfactual for each agent in the data using our model and parameters. We find that our metric does not map one-to-one to simple summary metrics like estimated risk aversion, productivity or past sales, emphasizing the need to carefully consider the joint impact of multidimensional types in such an evaluation, and the need for a formal model to compute such a metric.

Counter intuitively, we show that the most valuable agent to the firm is not necessarily the one with the highest productivity or profitability (defined as output − payout). Rather, we find the value of an agent is dependent on the distribution of types in the firm. For example, if there is only one "star" sales-agent in a firm and the rest are "duds," retaining the star will require the firm to provide high commissions, which is sub-optimal for incentivizing the remaining duds. Thus, it may be optimal to drop the star and to fine tune the commission to better align it to the characteristics of the duds. In this example, the star is not very valuable to the firm.[3] The model-based metric we compute incorporates this insight. More generally, we find in several situations that the value of an agent to the firm as a function of his productivity has an inverted "U-shape": intuitively, a low-productive agent is not of much value, but a highly productive one may also not be of much value because of the externality he induces on the actions of others via compensation. We use our metric to identify a set of agents who fall in bottom tail of the distribution of agents (least desirable from the firm's perspective). As a casual ex-post validation of this counterfactual exercise, the focal firm in our data actually ended up firing 4 workers who can be identified as not desirable ex ante according to the metric.

Finally, 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. Thus, the ability to choose agents seem to help balance the loss in incentives from the restriction to homogeneity, at least in this empirical example.

[3] Alternatively, it may be optimal to find another way outside of regular compensation to incentivize the "star" say, via fast-tracked promotions, public recognition, "President's Club" status, etc.

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 match between agents and 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's famous 2000a 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.'s 2007 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 depended 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 a particular ability. Unlike our set-up though, Lazear considers unidimensional agents in an environment with no asymmetric information or uncertainty.

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.

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. 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 our analysis.

2 The General Setup

Consider a pool of heterogeneous sales-agents indexed by $i = 1...N$ employed at an ongoing firm. The firm wishes to optimize the composition and compensation of its salesforce. Reflecting our empirical application, we assume the firm has divided its potential market into N geographic territories, and the maximum demand at the firm is for N sales-agents.[4] 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), and that the principal knows the agent's type (no learning or adverse selection).[5] Since sales are stochastic, the principal cannot back out the hidden effort from realized output, which generates the standard moral hazard problem.

In this paper, we abstract away from the sales-territory assignment problem (e.g., Skiera and Albers, 1998). For now, assume that the firm cannot replace the agents it fires (we discuss this in more detail below).

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,

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

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,

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

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

[5] In our data, learning about agent type is not of first-order importance 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 the adverse selection point later in the paper.

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 thus induces the chosen set of agents in $\mathcal{M}$ to stay and the rest to leave to pursue their outside option. This set-up is a reasonable way to capture the tradeoffs of a firm with an existing salesforce which is redesigning its salesforce contract so as to endogenously induce some agents to stay and some to quit. Modeling the problem for a new firm that is building a salesforce from scratch will require extending this model to formalize new agent hiring and learning, which this set-up abstracts away from.

Finally, we assume the firm has exclusive territories for its agent (consistent with our empirical context). To complete the model, we also need to specify what happens to demand from a territory managed by an agent if that agent is dropped from the firm. We use two assumptions, first that sales are 0 if a territory has no active agent, and the second, 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. The latter assumption 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.

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

$$\max_{\mathcal{M}\in\mathbb{M}_N} \Pi = \int \sum_{i\in\mathcal{M}} \left[S_i - \mathcal{W}_{\mathcal{M}}\left(S_i\right)\right] d\mathcal{F}\left(S_i|e_i\right) \tag{4}$$

with,

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

subject to the IC and IR constraints as before. Since $\mathcal{W}_{\mathcal{M}}$ is point-wise the optimal compensation plan for each sub-salesforce $\mathcal{M} \in \mathbb{M}_N$, 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

We now discuss the parametric assumptions we impose so as to operationalize the setup above for our empirical setting. We consider a firm that employs a pool of heteroge-

neous agents indicated by $i = 1, ..., N$. Each agent 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,

$$S_i = h_i + k_i e_i + \sigma_i \varepsilon_i \tag{5}$$

This functional has been used in the literature (see e.g. Lal and Srinivasan 1992; Holmstrom and Milgrom 1987) and interprets h as the expected sales in the absence of selling effort (i.e. $\mathbb{E}[S_i|e_i = 0] = h_i$) , k as the marginal productivity of effort and σ^2 as the uncertainty in the sales production process. As is usual, we assume that the firm only observes S_i and knows $\{h, k, \sigma\}$ for all agents. The density $\mathcal{F}(S_i|e_i)$ is induced by the density of ε_i. We will assume the firm sets linear contracts with compensation given as $\alpha_i + \beta S_i$ (see Holmstrom and Milgrom 1987 for some justifications of the optimality of linear contracts).

The agent's utility function is defined as CARA,

$$U_i = -\exp\{-r_i W_i\} \tag{6}$$

with wealth linear in output, and costs which are quadratic in effort,

$$W_i = \alpha_i + \beta S_i - \frac{d_i}{2} e_i^2 \tag{7}$$

The agent maximizes expected utility,

$$\begin{aligned} \mathbb{E}[U_i] &= -\int \exp\left\{-r\left(\alpha_i + \beta S_i - \frac{d_i}{2} e_i^2\right)\right\} d\mathcal{F}(\varepsilon_i) \\ &= -\exp\left\{-r\left(\alpha_i + \beta(h_i + k_i e_i) - \frac{d_i}{2} e_i^2 - \frac{r_i}{2}\beta^2 \sigma_i^2\right)\right\} \end{aligned} \tag{8}$$

The Certainty Equivalent is,

$$\mathcal{CE}_i = \alpha_i + \beta(h_i + k_i e_i) - \frac{d_i}{2} e_i^2 - \frac{r_i}{2}\beta^2 \sigma_i^2 \tag{9}$$

which implies that the optimal effort for the agent is,

$$e_i(\beta) = \frac{\beta k_i}{d_i} \tag{10}$$

3.1 The Principal's Problem

The principal treats agents as exchangeable and cares only about expected profits,

$$\mathbb{E}\left[\Pi\right] = \mathbb{E}\Big[\sum_{i=1}^{N}\left(S_i - \beta S_i - \alpha_i\right)\Big] \tag{11}$$

which is maximized subject to,

$$\begin{aligned} IC &: e_i\left(\beta\right) = \frac{\beta k_i}{d_i} \\ IR &: \mathcal{CE}_i \geq U_i^o \end{aligned} \tag{12}$$

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

$$\begin{aligned} \mathbb{E}\left[\Pi\right] &= \sum_{i=1}^{N}\left(\mathbb{E}\left(S_i\right) - \beta\mathbb{E}\left(S_i\right) - \alpha_i\right) \\ &= \sum_{i=1}^{N}\left(1-\beta\right)\left(h_i + k_i e_i\left(\beta\right)\right) - \alpha_i \end{aligned} \tag{13}$$

Further if the IR constraint is binding we have,

$$\alpha_i = U_i^o - \left[\beta\left(h_i + k_i e_i\right) - \frac{d_i}{2}e_i^2 - \frac{r_i}{2}\beta^2\sigma_i^2\right] \tag{14}$$

and substituting in we have,

$$\begin{aligned} \mathbb{E}\left[\Pi\right] &= \sum_{i=1}^{N}\Big[\ \left(1-\beta\right)\left(h_i + k_i e_i\left(\beta\right)\right) - U_i^o + \Big\{\beta\left(h_i + k_i e_i\left(\beta\right)\right) - \tfrac{d_i}{2}e_i\left(\beta\right)^2 - \tfrac{r_i}{2}\beta^2\sigma_i^2\Big\}\ \Big] \\ &= \sum_{i=1}^{N}\left[h_i + k_i e_i\left(\beta\right) - U_i^o - \frac{d_i}{2}e_i\left(\beta\right)^2 - \frac{r_i}{2}\beta^2\sigma_i^2\right] \end{aligned} \tag{15}$$

Differentiating with respect to β we get,

$$\frac{\partial\mathbb{E}\left[\Pi\right]}{\partial\beta} = \sum_{i=1}^{N} k_i e_i'\left(\beta\right) - d_i e_i'\left(\beta\right) - r_i\beta\sigma_i^2 \tag{16}$$

where,

$$e_i'\left(\beta\right) = \frac{\partial e_i\left(\beta\right)}{\partial\beta} = \frac{k_i}{d_i} \tag{17}$$

which gives the optimal uniform commission,

$$\beta^* = \frac{1}{1+\gamma} \tag{18}$$

where, γ,

$$\gamma = \frac{\sum_{i=1}^{N} r_i \sigma_i^2}{\sum_{i=1}^{N} \frac{k_i^2}{d_i}}$$

is an aggregate measure of the distribution of types within the firm. The salary, α_i^* can be obtained by substitution. Equation (18) 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 (18) 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 (18) 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 (15) evaluated at the optimal commission β^* as,

$$\mathbb{E}\left[\Pi\left(\beta^*\right)\right] = \sum_{i=1}^{N} \left(h_i - U_i^o\right) + \frac{\beta^*}{2}\left(\sum_{i=1}^{N} \frac{k_i^2}{d_i}\right) \tag{19}$$

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 (19) 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 (19) 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 (18), we can write the total payoff to

the principal with optimally chosen incentives as,

$$\mathbb{E}\left[\Pi\right] = \sum_{i=1}^{N}\left(h_i - U_i^o\right) + \frac{1}{2}\frac{\left(\sum_{i=1}^{N}\frac{k_i^2}{d_i}\right)^2}{\left(\sum_{i=1}^{N}\frac{k_i^2}{d_i} + \sum_{i=1}^{N} r_i\sigma_i^2\right)} \tag{20}$$

Equation (20) shows that at the optimal β^*, the payoff across agents is no longer separable across types. Equation (20) 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

For illustration, we choose the following agent profiles:

Parameter	**Agent 1**	**Agent 2**	**Agent 3**
r	1	2	1.5
σ	7	9	8
d	2	1.5	1
U^0	5	5	8
k	3	4	5
h	8	10	6

For what follows, note that agent 3 has the highest productivity (k), the middle level of risk aversion (r), and the lowest cost of effort (d) . As a base case, we first compute the optimal salary + commission for each agent separately, which we call the "fully heterogeneous" contract. Under this situation, the firm would retain each agent, and provide each a commission tailored to his type, setting a salary that leaves each his reservation utility. For the fully heterogeneous plan, we have:

Configuration	**Profits**	**Commission Rate**	**Sales**	**Effort**	**Compensation**
$\mathcal{M}$	$\mathbb{E}(\mathbf{\Pi}_{\mathcal{M}})$	β	$\mathbb{E}(\sum S_i)$	$\sum e_i$	$\mathbb{E}(W_i)$
$(1,1,1)$	9.10	$(0.0841, 0.0618, 0.2067)$	30.20	1.32	21.10

In the fully heterogenous plan, agent 3 gets the maximum commission (20.67%), agent 2 the least (6.18%), and the firm makes a profit of 9.1.

Now consider what happens when the firm is restricted to a common commission (but different salary) for each agent. We refer to this as the "partially homogenous" contract. The results are below. Solving for each configuration, we find that the firm would optimally drop agent 3 from the pool (expected payoff of \$8.51 with an optimal common commission rate to agents 1 and 2 of 6.71%). Including the agent in the pool requires the firm to set a higher commission rate (11.57%), which is too high for the other agents, reducing the firm's payoffs to \$8.32. Looking at the top four rows, we see that agent 3's presence in the pool exerts an externality on the others. If only agent 3 is retained, he would be paid a high-powered commission rate of 20.66%, while agents 1 and 2 prefer commissions of only 8.41% and 6.18% respectively. In this example, the firm is better off dropping the high-powered sales-agent from the pool so that it can incentivize the others to work harder: without agent 3, the firm can set an intermediate level of commissions that is better aligned with the other two types. The firm makes lower sales with only agents 1 and 2 (19.02 with only agents 1 and 2, versus 28.65 with all three), 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" type 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, emphasizing the need for a theory of behavior in order to assess the relative value of the sales-agents in the company.

Configuration	**Profits**	**Commission Rate**	**Sales**	**Effort**	**Compensation**
$\mathcal{M}$	$\mathbb{E}(\mathbf{\Pi}_{\mathcal{M}})$	β	$\mathbb{E}(\sum S_i)$	$\sum e_i$	$\mathbb{E}(W_i)$
$(0,0,1)$	0.58	0.2066	11.17	1.03	10.58
$(0,1,0)$	5.33	0.0618	10.66	0.16	5.33
$(0,1,1)$	5.17	0.1215	20.33	0.93	15.17
$(1,0,0)$	3.19	0.0841	8.38	0.13	5.19
$(1,0,1)$	3.49	0.1691	18.99	1.10	15.49
$(\mathbf{1,1,0})$	**8.51**	**0.0671**	**19.02**	**0.28**	**10.51**
$(1,1,1)$	8.32	0.1157	28.65	1.06	20.32

Finally, we provide a comparison to the "fully homogeneous" plan in which the firm sets the same salary and commission for every agent. The restriction to pay every agent the same salary reduces the profits for the firm significantly compared to the partially homogenous case. Intuitively, in this case only the reservation utility for the lowest type binds, and the other two types will obtain surplus above their reservation levels. The inability to extract this surplus by fine tuning salaries hurts the principal. The expected profits drop to \$1.99, a 76% decrease from the partially homogenous case (payoff of \$8.51). Not surprisingly, this case is rarely seen in practice, and is not of first-order real-world significance. It is however interesting to note that profits under the *partially* homogenous plan (\$8.51) come close to that of the fully heterogeneous plan (\$9.1), in spite of the restriction to common commissions. This suggests that the ability to pick agents is valuable and may compensate for the reduced incentives implied by the commonality in contract terms.

Configuration	**Profits**	**Commission Rate**	**Sales**	**Effort**	**Compensation**
$\mathcal{M}$	$\mathbb{E}(\mathbf{\Pi}_{\mathcal{M}})$	β	$\mathbb{E}(\sum S_i)$	$\sum e_i$	$\mathbb{E}(W_i)$
$(1,1,1)$	1.99	0.1165	28.67	1.06	26.69

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 3 had the highest productivity in terms of sales (11.17 vs. 10.66 and 8.38 for agents 1 and 2) and expended the highest stand-alone effort (1.03 vs .16 and .13 for agents 1 and 2). However, keeping the agent in the pool severely 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 a new Boolean optimization scheme that allows us to obtain the optimal salesforce composition without complete enumeration.

4 Cross Entropy Approach

The simple example presented above used a complete enumeration of possible salesforce compositions to examine profitability. Real-world sales-forces often numbers in the hundreds or thousands, and this approach is not practical. Finding the optimal configuration of agents is a large-scale integer-programming problem. To operationalize the theory to real-world settings, we experimented with several integer-programming algorithms and found significant success using the "Cross-Entropy" algorithm. Cross-entropy is a relatively new

optimization technique introduced first by Rubenstein (1997) in the context of sampling rare events. The approach has recently been adapted to combinatorial optimization and pseudo-Boolean optimization (De Boer et al. 2005). Since application of cross-entropy is new in the context of salesforce settings, we provide a short overview below. Readers who are uninterested in the technical details of the optimization can skip this section to the next, which discusses the empirical application.

The idea of the cross-entropy approach is to choose a parametric density (a product of Bernoullis for our context) that generates candidate solutions. At each iteration a set of candidates are "scored" on the basis of their profitability and the top ρ-quartile subset of solutions is retained. This "elite" subset is then used to update the parametric proposal density. This update entails using maximum likelihood to find parameters for the candidate density that rationalize the elite sample. The algorithm then iterates until convergence. The properties of the algorithm and the rates of convergence are discussed in Margolin (2004) and Costa et al. (2005). Please see those papers for further details.

The Algorithm

1. Initialize $\mathbf{p}_0, R, T, \rho$ and $\{\alpha_t\}_{t=1}^{\infty}$
2. At each iteration t:
   (a) Generate a set of configurations $\mathbf{M}_t^{(r)}, r = 1..R$ from $f\left(\mathbf{M}|\mathbf{p}_{t-1}\right)$
   (b) Compute $\Pi^{(r)} = \mathbb{E}\left(\Pi\left(\mathbf{M}_t^{(r)}\right)\right)$ for all r and order $\Pi^{(1)} \leq \Pi^{(2)} \leq ... \leq \Pi^{(R)}$
   (c) Let $R^{\rho} = \lceil (1-\rho) R \rceil$ and compute the $(1-\rho)$-quantile of profits: $\hat{\gamma}_t = \Pi^{(N-N^{\rho}+1)}$
   (d) Let $\mathcal{G}_t$ denote the set of indices r such that $\Pi_t^{(r)} \geq \hat{\gamma}_t$, and let $R_G = |\mathcal{G}_t|$
3. For each $i = 1..N$ calculate
$$\omega_{it} = \frac{1}{R_G} \sum_{r \in \mathcal{G}_t} \mathbf{1}_{\left(\mathbf{M}_{it}^{(r)}=1\right)}$$
4. Update
$$p_{it} = (1-\alpha_t)\, p_{it-1} + \alpha_t \omega_{it}$$
5. If $t = T$, or if
$$\max_{1 \leq j \leq N} \{\min\{p_{it}, 1-p_{it}\}\} \leq \varepsilon$$
for small ε, then stop, or else set $t = t+1$ and return to step 2.

Consider the 3-agent example we discussed earlier − we wish to maximize expected profits by choosing the optimal combination of agents and the corresponding optimal compensation contract. To start, we use a product of three (independent) Bernoulli distributions

(with initial parameter $p = 0.05$ for each) to generate candidate constellations. At the initial parameter value, each agent has a equal probability of being included or excluded. Suppose we generate 20 candidate constellations at each iteration. We compute the optimal contract and the expected profits for each of these possible 20 constellations. The constellations are then ranked on the basis of the expected profits, and the top ρ solutions are used to update the parameters of the Bernoulli distributions. We move on to the next iteration. This process continues till the candidate density converges. When it does, the constellation at the convergence is denoted the optimal configuration.

A few points are worth noting. First, the updating step described in the algorithm is tailored to the specifics of our application. Other applications might require tuning the candidate density, the update rule, the number of draws and/or the size of the elite sample. These affect the quality of the results and the rate of convergence. For example, if we started with parameter values for probabilities close to zero or one, the algorithm would take an inordinate amount of time to find its way to the maximum. Similarly, large elite groups do not help, while ones that are too small make the algorithm "jumpy." These elements and issues are found in most optimization algorithms in some form or another. We simply wish to point out that the cross-entropy approach is not free from the need for human management and tuning. All results presented below are based on the cross-entropy approach. In using alternative competing methods (particularly Simulated Annealing and Genetic Algorithms) in simulations, we found the cross-entropy method superior. To assess the robustness of the results presented herein we also attempted to find "better" solutions by taking the cross-entropy results as starting points in the other more accepted, approaches. In no case did the alternate approaches find a better solution to that obtained by the cross-entropy algorithm.

5 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). 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. We note in passing that prices 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 is also characterized by significant new product introduction 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 competitors. While agent's frequency of visiting doctors is monitored by the firm, the extent to which he "sells" 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).[6]

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 N 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

[6]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.

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 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,

$$u_t = u(Q_t, a_t, I_t, \chi_t = 1) = \mathbb{E}[W_t] - r \times \text{var}[W_t] - \frac{de_t^2}{2} \tag{21}$$

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 (21) collapses to exactly the form denoted in equation (9) for the certainty equivalent. We can thus interpret equation (21) 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,

$$u_t = u(Q_t, a_t, I_t, \chi_t = 0) = U^0 \tag{22}$$

In this model, sales are assumed to be generated as a function of the agent's effort, which is choose 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.,

$$Q_{t+1} = \begin{cases} Q_t + q_t & if \quad I_t < N \\ 0 & if \quad I_t = N \end{cases} \tag{23}$$

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,

$$a_{t+1} = \begin{cases} a_t & if \quad I_t < N \\ \sum_{k=1}^{K} \theta_k \Gamma \left(a_t, Q_t + q_t \right) + v_{t+1} & if \quad I_t = N \end{cases} \tag{24}$$

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. Finally, the transition of the third state variable, months since the beginning of the quarter, is deterministic,

$$I_{t+1} = \begin{cases} I_t + 1 & if \quad I_t < N \\ 1 & if \quad I_t = N \end{cases} \tag{25}$$

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 $I_t < N$, the agent's present-discounted utility under the optimal effort policy can be represented by a value function that satisfies the following Bellman equation (see Misra and Nair 2011),

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

$$\max_{\chi_{t+1} \in (0,1), e>0} \left\{ \begin{array}{l} u(Q_t, a_t, I_t, \chi_t, e; \Omega, \Psi) \\ +\rho \int_\varepsilon V(Q_{t+1} = Q(Q_t, q(\varepsilon_t, e)), a_{t+1} = a_t, I_t + 1, \chi_{t+1}; \Omega, \Psi) f(\varepsilon_t) d\varepsilon_t \end{array} \right\} \tag{26}$$

Similarly, the Bellman equation determining effort in the last period of the sales-cycle is,

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

$$\max_{\chi_{t+1} \in (0,1), e>0} \left\{ \begin{array}{l} u(Q_t, a_t, N, \chi_t, e; \Omega, \Psi) \\ +\rho \int_v \int_\varepsilon V(Q_{t+1} = 0, a_{t+1} = a(Q_t, q(\varepsilon_t, e), a_t, v_{t+1}), 1, \chi_{t+1}; \Omega, \Psi) \\ \qquad \times f(\varepsilon_t) \phi(v_{t+1}) d\varepsilon_t dv_{t+1} \end{array} \right\} \tag{27}$$

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

maximizes the value function,

$$e\left(\mathbf{s}_t;\Omega,\Psi\right)=\arg\max_{e>0}\left\{V\left(\mathbf{s}_t;\Omega,\Psi\right)\right\} \tag{28}$$

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

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

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 $e^*\left(\mathbf{s}_t;\Omega\right)$, given a guess of the parameters Ω. Because $\widehat{e}_t$ is known, we can then use $\widehat{e}_t=e^*\left(\mathbf{s}_t;\Omega\right)$ 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 any cross-agent restrictions, thereby obtaining a semi-parametric distribution of the types in the firm.

The question remains how the effort policy, $\widehat{e}_t=\hat{e}\left(\mathbf{s}_t\right)$ 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.[7] For the counterfactuals in this paper, we need estimates of $\left\{h,k,d,r,U^0,\mathcal{G}_\varepsilon\left(.\right)\right\}$. Here, we assume that $\mathcal{G}_\varepsilon\left(.\right)\sim N\left(0,\sigma^2\right)$. So, we need $\left\{h,k,d,r,U^0,\sigma\right\}$. 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}. 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 in-

[7] See also, Steenburgh (2008) and Larkin (2010) who note that effort is a function of how far away the agent is from his quota.

corporation 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 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. Since we are using data from the new plan, 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 cross-entropy approach 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 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 was to be left empty, sales would still accrue on account of the brand or because the firm might use some other (less efficient) selling approach. 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 six 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 six agents who are dropped 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 dropped from the firm under 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 eliminated agents not uniformly at the bottom of the heap in terms of expected sales or profit contribution under the fully heterogeneous plan. For example, Agent #33, one of the agents who were dropped, had 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 was 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). Eliminating this agent allows the firm to improve the contract terms of other agents thereby increasing

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

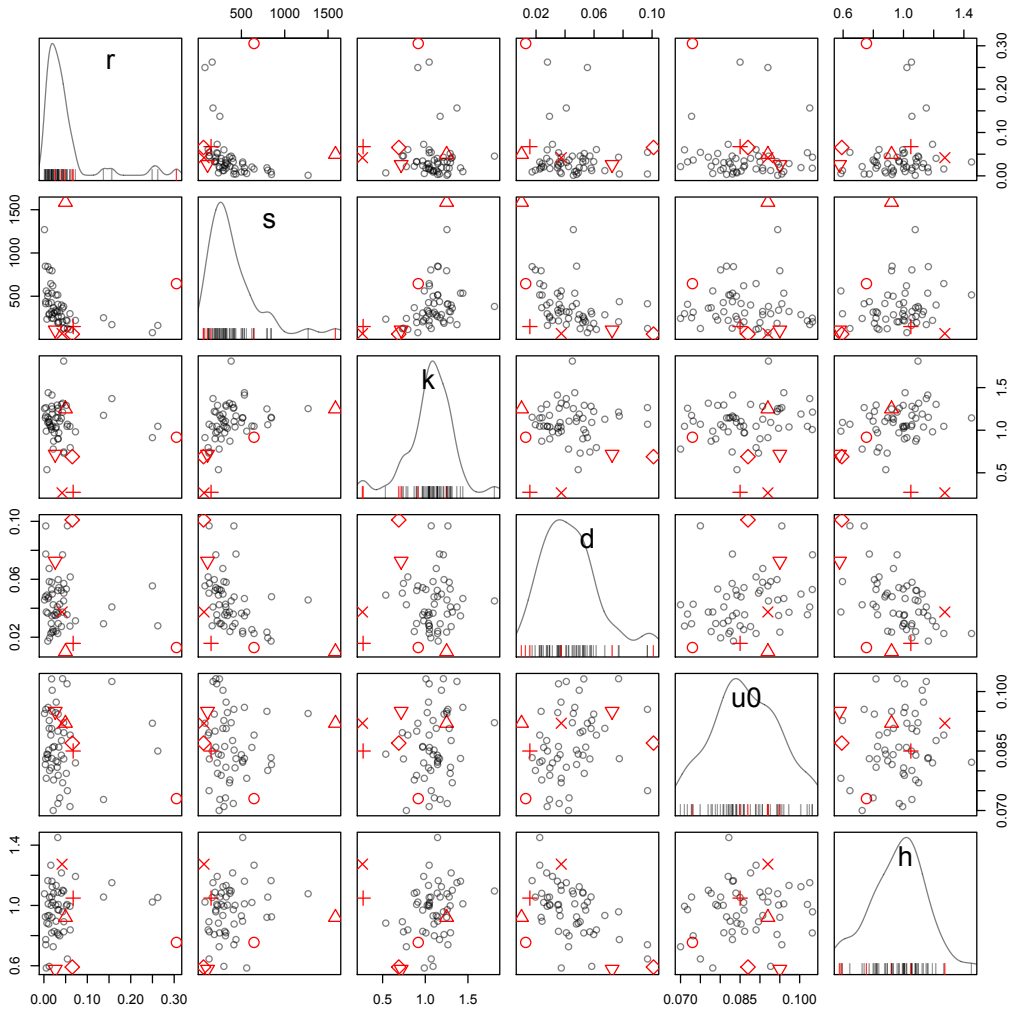
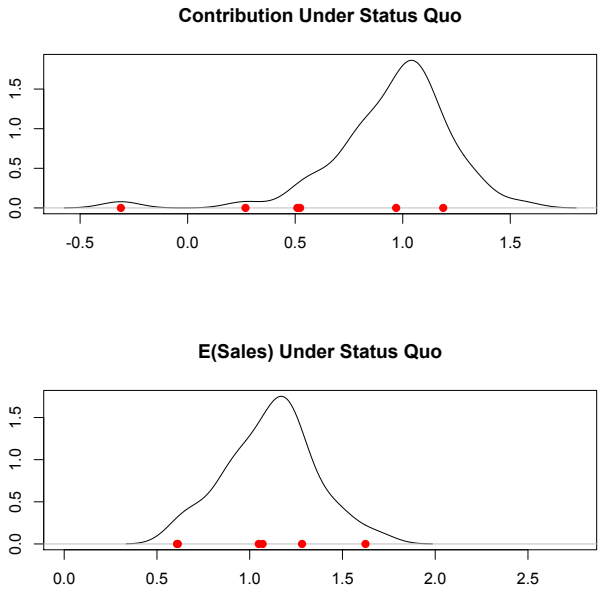


Figure 2: Profitability and Sales of Eliminated Sales Agents



profits.[8] Other agents were similarly eliminated on account of some other externality that impacted the compensation contract.

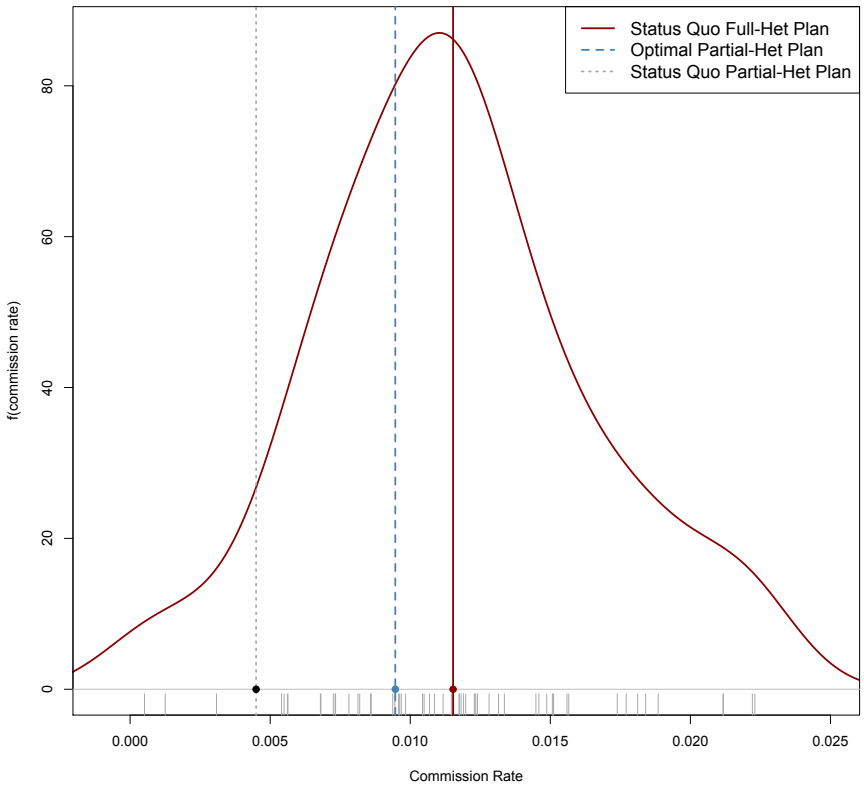
Figure 1 and 2 accentuates the difficulties of ranking agents as desirable or not on the basis of a single type-based metric. In a subsequent section below, we discuss a theory-based metric that makes sense for assessing agents in such a setting.

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%). Under

[8] More generally, our point is not that this agent should necessarily be fired as a matter of policy, but that in the absence of firing this agent, the firm should find another way outside of salary + commission to incentivize the agent.

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



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 with-

	Composition →	
Compensation ↓	*Status Quo*	*Optimal*
Fully Heterogeneous	\$60.56MM	\$60.56MM
Partially Homogenous	\$55.78MM	\$59.14MM

Table 1: Profits under Fully and Partially Heterogenous Plans

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

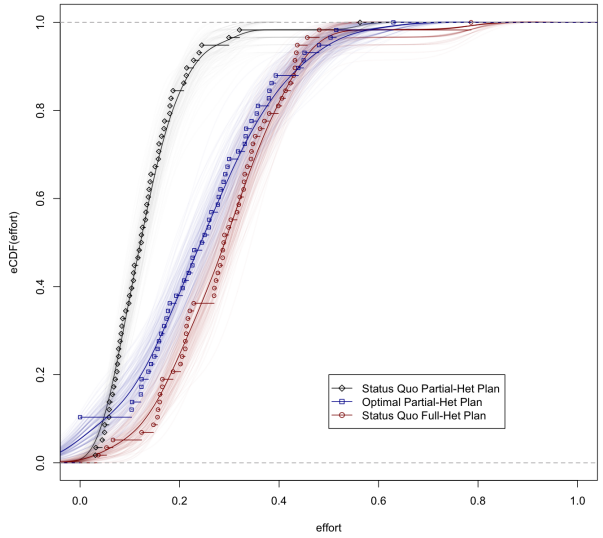
In contrast to the fully heterogeneous compensation structure, there is a significant difference in profit levels when compensation plans cannot be customized. Looking at the above table, partially homogenous plans with the ability to fine-tune composition come very close to the fully-heterogeneous plan in terms of profitability (\$59.14MM 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 \$3.36MM, 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.

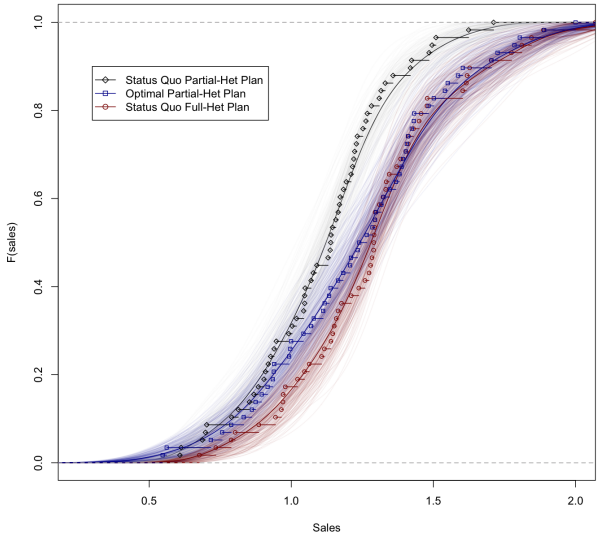
We now assess the extent to which profits at the *individual sales-agent* level under the partially homogenous plan combined with the ability to choose the composition of agents, approximates the profitability under the fully heterogeneous plan (the baseline or best-case scenario). In Figure 6, we plot 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.

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



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.

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). 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. This intuition shows that, all things equal, the firm constrained to a uniform contract prefers the heterogeneity across agents is minimal.

Thus, when given the option to choose both agents and plans, the firm finds a subset of relatively homogenous agents who can be incentivized close to optimally. Thus, heterogeneity reduction is the root of the mechanism that generates the improved profitability. Put differently, note that 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 the 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,[9]

$$d_{\mathcal{M}} = \text{tr}\left(\Sigma_{\mathcal{M}}\right) \tag{29}$$

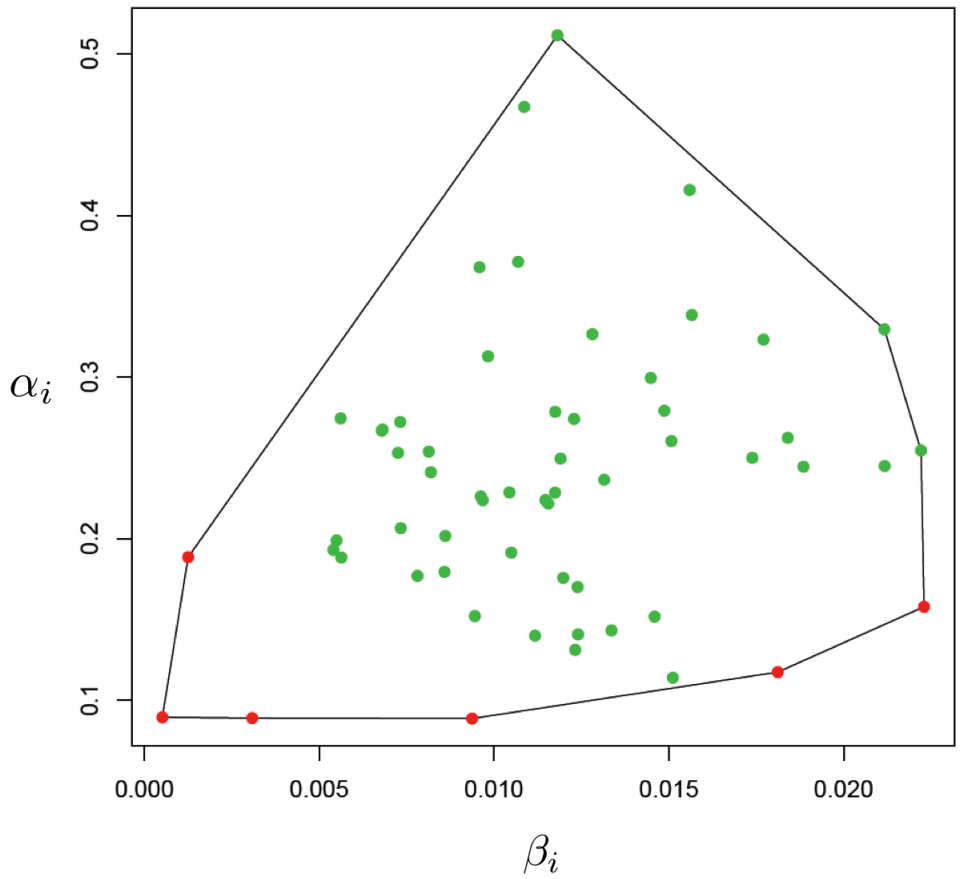
We find that $d_{\mathcal{M}_{\text{StatusQuo}}} = 78,891.6$, and $d_{\mathcal{M}_{\text{Optimal}}} = 51,921.8$, where $d_{\mathcal{M}_{\text{StatusQuo}}}$ is the trace under the optimally chosen partially homogenous plan while retaining all agents in the firm, and $d_{\mathcal{M}_{\text{Optimal}}}$ 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 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

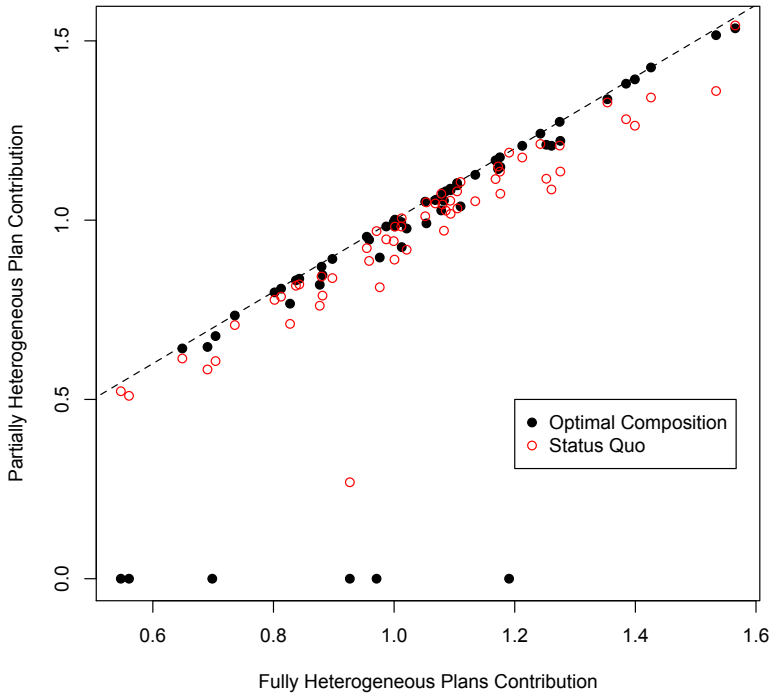
[9]The trace of a matrix is the sum of its diagonals.

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.

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



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 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,

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

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

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

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.

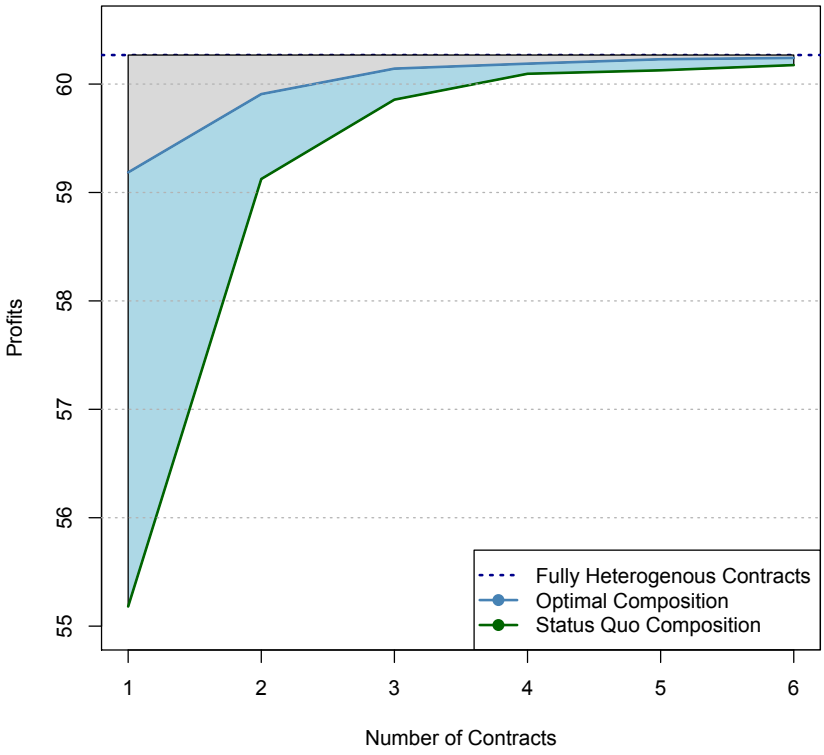
We now discuss the situation where we allow the firm to sort agents into divisions and to optimize its composition. We see the results are similar to the previous case, but the firm is able to achieve a higher level of profit gains with fewer divisions (the blue-line is always above the green-line). 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 profitabaility with sorting. In Figure (8) we plot the log-distortion implied by the divisions against the number of contracts offered.[10] The log-distortion is simply a summary measure of the average heterogeneity within a division. 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

[10]We define distortion as the mean squared deviation from the average salesperson (c_X).

$$d_{\mathcal{M}} = \frac{1}{|\mathcal{J}|} (X - c_X)' (X - c_X)$$

Figure 7: Performance of Divisions in Managing Heterogeneity

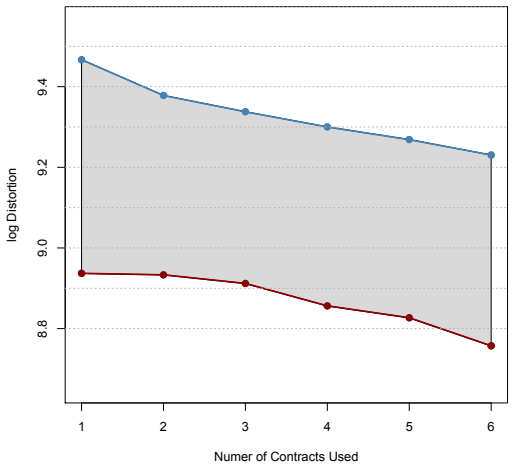


use the increased flexibility to better compensate those that are retained. As a result the heterogeneity in the retained agent pool is managed much better. Ultimately, as the number of contracts increase to match the total number of agents, the two curves would coincide. That is, if every agent got 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.

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 hetero-

Figure 8: Heterogeneity Reduction in the Saleforce with Many Divisions



geneity 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 The Value of an Agent

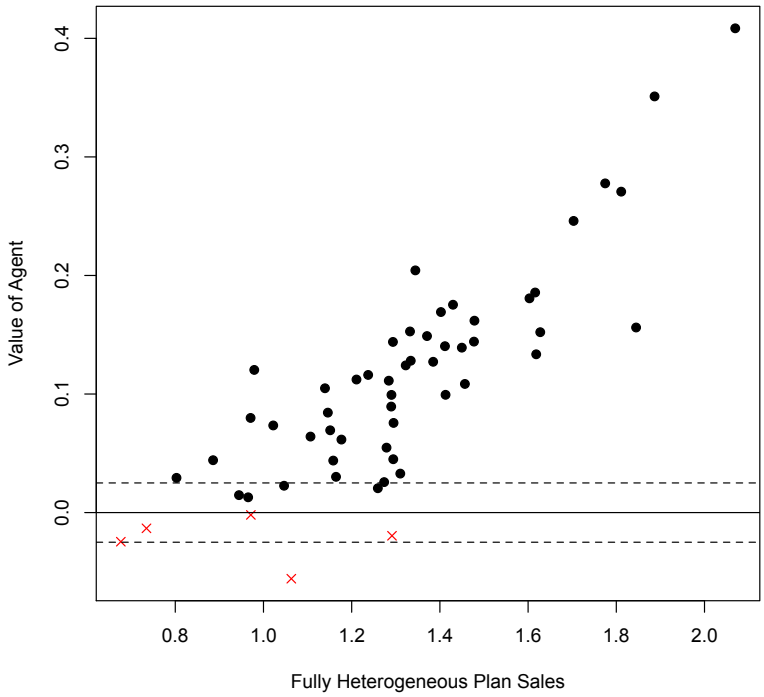
To close the paper, we use the model above for a salesforce assessment exercise that complements the human resources management of the sales and marketing organization in the firm. Our goal is to compute a metric that encapsulates the value of each agent to the firm. In addition to incremental contribution, our metric incorporates the contractual externality imposed by an agent implied by the restriction to uniform commissions. The main insight behind our metric is that the value of the agent is not simply the present-discounted value of the sales minus payouts to the agent. Rather, the value of the agent has to be assessed relative to a counterfactual world, in which the firm re-optimizes both the composition and the compensation of the agent pool that remains with the firm once that agent departs. Essentially, the loss of an agent changes the distribution of types within the firm. The change in the type distribution necessitates a change to the way the remaining agents are incentivized, which requires re-adjusting both the contract structure and the salesforce structure under the new situation. A full assessment of the value of an agent will take this counterfactual into account.

We compute this metric as follows. Starting with the first agent in the pool, we compute what would be the profits to the firm if that agent were surely retained in the salesforce, but the firm were to optimize both compensation and composition to the entire salesforce under the restriction that the agent always stays. We store these numbers in memory. We then drop agents sequentially from the firm. When each agent is dropped, we use our model and estimates to compute what would be the profits to the firm under the optimal composition and composition of the remaining agent pool without the focal agent. We compute our metric of an agent's value as the total profit to the firm with the agent in the firm minus the total profit to the firm without the agent, wherein, in both scenarios, we allow the firm to re-optimize the salesforce and its incentives as described above. We repeat this exercise for each agent in the firm, simulating the profit differences for each. Each simulation is a separate counterfactual pair. Formally, define $\Pi\left(\mathbb{M}_N\right)$ as the profit to the principal when choosing composition $\mathcal{M}$ and compensation $\mathcal{W}$ optimally from set $\mathbb{M}_N$,

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

Let the set of agent-configurations in which agent i is included be denoted $\mathbb{M}_N^{(i)}$, and those

Figure 9: Value Metric Plotted Against Agent's Profitability Under the Fully Heterogeneous Plan



in which he is not be denoted $\mathbb{M}_N^{(-i)}$. Then, our value metric is defined as,

$$\mathcal{V}_i = \Pi\left(\mathbb{M}_N^{(i)}\right) - \Pi\left(\mathbb{M}_N^{(-i)}\right)$$

This value metric has the advantage that it retains a link to the theory and that it summarizes the effect of the multivariate nature of each agent's type on his attractiveness to the firm.

Figure (9) plots the value metric computed across agents (y-axis). For ease of comparison to contribution-based metrics, we also plot each agent's profitability under the fully heterogeneous plan on the x-axis. The agents who are dropped under the optimal composition are denoted by "x"-s (one agent is very undesirable and has a large negative value metric; he is not included in the plot for pictorial clarity). Figure (9) documents the wide heterogeneity in the salesforce: some agents are very valuable to the firm compared to the average. Further, there are several whose computed value is negative. Not surpris-

ingly, these are the ones who are dropped from the firm in the optimal configuration.[11] While agents with negative value are identified by the metric, a useful by-product is that it also identifies a set of agents who are close to the zero cutoff. To the extent that there is statistical noise in the metric, this pin-points a set of agents the firm should monitor closely or place "on probation". In this sense, the value metric is a useful statistic that complements the human resource management of the sales and marketing function.

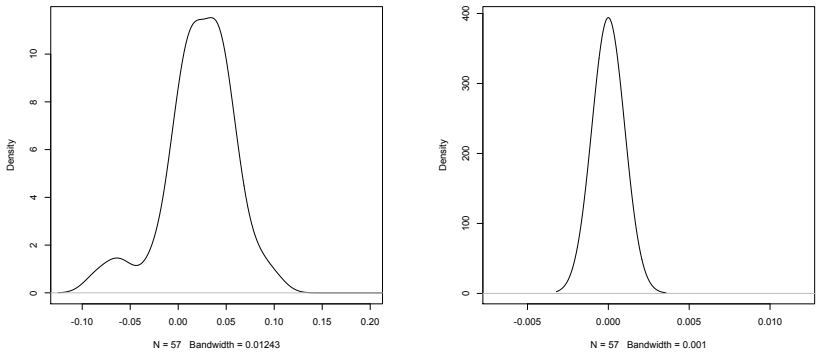
Another take away from Figure (9) is that profitability is a useful, but not fully diagnostic metric for assessing the quality of agents. For instance, there are several agents close to the median level of profitability under the fully heterogeneous plan who are nevertheless in the band for probation or are altogether dropped from the firm. These agents are picked out for management's attention by our metric. The pictures look similar if we plot the value metric against expected sales under the fully heterogeneous plan (instead of profits), or against alternative data-based measures like summaries of observed past sales/profits on the x-axis.

Finally, Figure (9) assesses the effect on *total* profits when an agent is dropped from the pool, compared to when he's retained. The advantage of our approach is that we can also simulate out the effect on *every other* agent when a given agent is dropped. To illustrate, we plot the density across agents of the difference in profitability when a focal agent is dropped from the firm versus when he's retained. We see that agent 12 (who is dropped from the firm in the optimal configuration) induces a negative externality on a large proportion of the agent pool: the profitability of most other agents rise when #12 is dropped. Figure (10) also plots how the profitability of agents change when another agent (#21), who is retained in the firm under the optimal configuration, is instead dropped from the firm. Dropping #21 does improve the profitability of some agents, but the net effect for the firm is positive, and he is thus retained. The advantage of our structural approach is that it facilitates such detailed assessments by retaining a link to theory.

Discussion We close the paper with a discussion of some additional implications of the above approach to value. The logic behind the value metric reveals a sense in which firing and retaining sales-agents involve economies of scope. In particular, the value of an agent A when contemplating firing only agent A, is different from the value of A when contemplating firing *both* agents A and B simultaneously. The joint departure of A and B together from the firm would imply a different profit outcome for the firm via changed incentives and composition. In essence, both these are two different counterfactuals and therefore imply two different valuations. This generate several interesting implications. As a practical example, this implies for instance, that the value of an agent is different when

[11] It is not necessary (but is likely to be the case typically) that an agent with a negative value is always dropped from the firm in the optimal configuration - it happens to be the case in this example. The intuition as discussed further in this section is that the firm's decision to drop other agents simultaneously affects a focal agent's valuation.

Figure 10: Density across Agents of Difference in Profitability when the Focal Agent is Dropped from the Firm versus Retained



(a) Effect of dropping Agent # 12 (b) Effect of dropping Agent # 21

the firm is thinking of letting him go as part of bad performance (a one-off firing event) versus as part of a company-wide layoff or downsizing (when *many* agents will be let go simultaneously). Another immediate implication is that the value of a group of agents can be very different than the sum of their values. Thus, it may be that a set of agents may not look very attractive in isolation, but may be very attractive jointly. This may drive the firm to do "coordinated hiring." The reverse may also be true: a set of agents may be attractive individually, but may be unattractive jointly, nudging the firm towards a coordinated firing policy.

Finally, there is a sense in which the model predicts a "run" in firing, wherein firing a few agents leads to one or more agents to also be fired. To see this, note than its possible that firing agent A (or many agents like A) induces the firm to re-optimize its composition, and in that re-optimized composition, agent B is let go, even though agent B would not be fired from the firm if A has stayed with the company. In this sense, A imposes another kind of externality on other salespeople. If A (or many agents like A) leave the company due to a random shock (A's family moves elsewhere), it may be that it then becomes optimal for the firm to fire B!

This also exposes a hidden danger (or advantage, as the case may be) of uniform contracts. In the example above, if B anticipates that A's departure would harm him via the changed compensation and composition implied by his departure, it may be optimal for B to offer a trade to A to induce him to stay. This may lead to "trading favors" or pro-social behavior that may help the firm and improve "culture". On the other hand, one can imagine a situation where B anticipates that A's departure is beneficial to him (via increased commissions or a more preferred remnant sales-pool). In this case, B has an incentive to take actions to get A fired. In this example, uniformity results in harmful

incentives that may hurt the firm and debilitate internal "culture". Exploring these issues more formally is beyond the scope of this paper, and we leave this for subsequent research.

More broadly then, our high-level point is that valuation of sales-agents attains meaning only relative to a counterfactual. Unless the counterfactual is clearly stated, the valuation exercise is imprecise. Our model and approach facilitates these counterfactuals when stated.

7 Conclusion

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. We present a metric that can value agents in the firm incorporating this effect. 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. 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. 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. 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.

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

5. 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.

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

7. Costa, A., O. D. Jones, D. Kroese. (2007). "Convergence Properties of the Cross-Entropy Method for Discrete Optimization," Operations Research Letters, 35:5, 73-580.

8. De Boer, P-T., Kroese, D.P, Mannor, S. and Rubinstein, R.Y. (2005). A Tutorial on the Cross-Entropy Method. Annals of Operations Research, 134 (1), 19-67.

9. 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.

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

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

12. 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.

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. Holmstrom, B. and Milgrom, P. (1994). "The Firm as an Incentive System," American Economic Review, 84(4), 972-991.

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

17. 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.

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

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 Sales-forces: 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. Margolin, L. (2004). "On the Convergence of the Cross-entropy Method," Annals of Operations Research, 134, pp. 201–214.

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

33. 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.

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

35. 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

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

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

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

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

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

41. Rubinstein, R.Y. (1997). "Optimization of Computer simulation Models with Rare Events," European Journal of Operations Research, 99, 89-112.

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

43. Steenburgh, T. (2008). "Effort or Timing: The Effect of Lump-sum Bonuses," Quantitative Marketing and Economics, 6:235256.

44. Zoltners, A., P. Sinha and G. Zoltners. (2001). "The Complete Guide to Accelerating Sales-force Performance," American Management Association, New York.