The Journal of FINANCE
The Journal of THE AMERICAN FINANCE ASSOCIATION

THE JOURNAL OF FINANCE • VOL. LXXX, NO. 2 • APRIL 2025

Worker Runs

FLORIAN HOFFMANN and VLADIMIR VLADIMIROV*

ABSTRACT

The voluntary departure of hard-to-replace skilled workers worsens firm prospects, which can lead to additional departures. We develop a model in which firms design compensation to limit the risk of such "worker runs." To achieve cost-efficient retention, firms combine fixed wages with dilutable compensation—such as vesting equity or bonus pools—which pays remaining workers more when others leave but gets diluted otherwise. Compensating (identical) workers with differently structured compensation, that is, with a different mix of output-dependent and output-independent pay, can further mitigate the risk of worker runs by ensuring a critical retention level in a cost-efficient way.

IN JUNE 2021, THE FINANCIAL press reported extensively on Credit Suisse's fight to stem the exodus of senior and junior employees across multiple divisions. The trigger for these departures was the bank's exposure to the spectacular collapses of Archegos and Greensill Capital, which dented its profits. Senior bankers who were not involved in these affairs, and whose divisions would have otherwise led the bank to a record quarterly profit, were reportedly furious, prompting them to leave. As a result, bankers began departing in droves, as "nobody wanted to be the last man standing." At

*Florian Hoffmann is at KU Leuven. Vladimir Vladimirov is at University of Amsterdam and CEPR. This paper benefited significantly from comments by the co-Editor Philip Bond, an anonymous Associate Editor, and two anonymous referees. We further thank Dan Bernhardt, Arnoud Boot, Matthieu Bouvard, Tolga Caskurlu, Alvin Chen, Daniel Ferreira, Torsten Jochem, Tomislav Ladika, Andrey Malenko, Rich Mathews (Cambridge discussant), Yuliyan Mitkov, Stefan Obernberger, Enrico Perotti, Konrad Raff (EFA discussant), Elena Simintzi, Martin Szydlowski, Spyros Terovitis, Yenan Wang, Ed van Wesep, Liyan Yang (WFA discussant), Yue Yuan, and conference and seminar participants at the Western Finance Association (Huntington Beach), European Finance Association (Helsinki), European Economic Association (Cologne), Cambridge Corporate Finance Theory Symposium, Finance Theory Group (Stockholm), Toulouse School of Economics, University of Amsterdam, University of Bristol, and University of Warwick for constructive feedback. We have read *The Journal of Finance* disclosure policy and have no conflicts of interest to disclose.

Correspondence: Vladimir Vladimirov, Finance Group, University of Amsterdam, The Netherlands; e-mail: vladimirov@uva.nl.

DOI: 10.1111/jofi.13424

the time, pundits worried that the snowball effect of such departures could seriously erode dealmaking and the bank's market position.[1]

The collective turnover problem faced by Credit Suisse illustrates a broader phenomenon recognized by both academics and industry practitioners (Felps et al. (2009), Hausknecht and Trevor (2011), Heavey, Holwerda, and Hausknecht (2013), Hancock et al. (2013)). A recent industry study, which focused on quantifying contagious turnover in a sample of more than 10 million workers, documents that an initial resignation in a team of workers is associated with a 9% higher probability of further resignations within the same team Derler et al. (2023). Such contagion effects have arguably contributed to the fact that over the last 20 years, 23% of workers have quit their jobs each year (Gittleman (2022)), which is associated with staggering costs for firms. In the United States, the annual cost related to replacement, training, and lost productivity is reported at over a trillion dollars (McFeely and Wigert (2019), Barnette et al. (2024)), and Credit Suisse alone estimated in 2014 that reducing turnover could save it $100 million per year. Some of these costs are inevitable (and unrelated to contagion), as employee turnover is part of a natural and efficient process of workers finding a better match for their skills. However, it is in firms' interest to lower the part of these substantial costs that is due to inefficient contagious turnover, that is, the turnover resulting from a coordination failure among workers in fundamentally sound firms.

importance of the study: start with phenomenon

this is the focus

In this paper, we develop a novel model of collective turnover and show how compensation design that explicitly considers the dependence of workers' decisions to leave on those of their coworkers can achieve cost-efficient collective retention.[2] We identify and compare three compensation design solutions to the problem of contagious turnover: (i) fixed compensation; (ii) "dilutable" compensation, which promises higher pay if other workers leave but gets diluted if more workers stay and includes deferred equity-based pay or bonus pools as examples; and (iii) asymmetric compensation, whereby firms offer ex-ante identical workers compensation packages that can differ in size (level of pay) and structure (ratio of output-dependent to output-independent pay). We show how these versatile solutions are optimally combined to achieve cost-efficient collective retention in a variety of situations, allowing for realistic resource constraints and (information) frictions.

Worker runs. Our main modeling innovation lies in capturing the strategic complementarities in workers' decisions to leave, that is, the problem that the departure of some workers or the risk of such departures makes other workers more likely to leave even fundamentally healthy firms. In our model, the reason that workers may want to leave is that after being hired, they learn about the firm's prospects by privately observing shocks to its productivity.

[1] See "Credit Suisse fights to stem exodus as top US dealmaker quits" June 17, 2021, *Financial Times*.

[2] Our focus on compensation is motivated by evidence that compensation is both a primary reason for voluntary quits and the most important retention tool (Erickson et al. (2020), Breitling et al. (2021)).

These shocks affect workers' beliefs about how their on-the-job pay compares to available outside options and may lead some workers to leave. Crucially, the departure of hard-to-replace workers reduces firm productivity and thus the value of remaining workers' output-dependent compensation, such as equity or bonuses. This may trigger remaining workers to follow suit, resulting in a "worker run." Worker runs are both more likely and more costly in firms that rely heavily on teams of hard-to-replace workers with complementary skills. Typical examples are startups whose success depends crucially on retaining well-functioning teams consisting of key personnel in research and development, management, sales, and marketing. Other examples include consulting, advisory, investment banking, law firms, and private equity partnerships. Worker runs also occur in large firms organized around teams (such as Credit Suisse), with turnover contagion being two to three times higher in smaller teams (Derler et al. (2023)).

In our model, the objective of optimal compensation design is to achieve collective retention at the lowest compensation costs. This requires: (i) mitigating the strategic complementarities in workers' decisions to leave, which is our paper's primary innovation, while (ii) minimizing workers' information rent arising from their private information about how their on-the-job pay compares to their outside options. Achieving the first objective requires smoothing workers' expected compensation over different retention scenarios. This can be simply achieved with deferred fixed compensation that pays the same in all cash-flow states, as the value of fixed compensation is independent of the retention of other workers.[3] However, offering high enough fixed pay to guarantee retention may result in paying workers much more than their outside options in many states of the world, resulting in large rents for workers. Moreover, resource constraints may limit the amount of fixed pay that firms can offer. In these cases, the firm offers workers output-dependent compensation such as equity-based pay or performance bonuses. We show that whenever the firm uses such compensation, it can lower the cost of retaining workers by making compensation dilutable or offering asymmetric compensation.

Dilutable compensation. To counteract the decrease in the expected value of output-dependent compensation when productive employees leave the firm, the firm can offer "dilutable" compensation that promises remaining workers higher pay when other workers are leaving and dilutes workers' pay when retention levels are high. Examples of such "dilutable" compensation include paying workers with time-vesting equity (or options) or offering profit-sharing bonuses. Consider time-vesting equity: the standard argument is that such compensation boosts retention by deferring workers' pay, but other types of deferred pay arguably have the same effect. Our novel insight is that time-vesting equity further helps by tying workers' compensation to the firm's retention level. To illustrate, suppose a firm with 100 shares outstanding sets aside 100 additional shares for compensation purposes and promises each of its two workers that they will receive 50 time-vesting shares if they stay

[3] To improve retention, all compensation is optimally deferred and forfeited if the worker leaves.

sufficiently long with the firm. If both workers stay, each worker will own 50/200 = 25% of the firm's outstanding equity. However, if one worker leaves and forfeits their equity compensation, the remaining worker's equity stake will increase to 50/150 = 33%. Thus, even though the firm's equity value may drop when the first worker leaves, the higher percentage of equity ownership counteracts the second worker's incentive to leave. Equally important is that if the firm manages to retain both workers, the higher retention level dilutes each worker's percentage equity ownership, allowing the firm to retain the two workers at lower cost. Notably, dilution is not restricted to equity-based pay, and it can be easily decoupled from firm size. For example, firms may offer profit-sharing bonus pools at the division or team level or retention bonuses, which can also make contracts dilutable.

The optimal degree of dilution in workers' compensation is higher when firm output is more sensitive to retention and workers are paid with a higher share of output-dependent pay. In our model, the share of output-dependent compensation is determined by the firm's objective of minimizing workers' information rent by matching their on-the-job pay as closely as possible to their outside options subject to possible resource constraints. This objective calls for output-dependent pay if the firm's prospects positively correlate with workers' outside options (Oyer (2004)). In this case, output-dependent pay ensures that on-the-job pay increases following positive (systematic) shocks, guaranteeing retention when workers' outside options are high, and decreases following negative (systematic) shocks, lowering workers' rents when their outside options are worth little. Taken together, firms optimally offer more output-dependent—and therefore more dilutable—compensation if they are more cash-constrained or if workers' outside options are more sensitive to systematic shocks than firm output.

Asymmetric compensation. Another way to reduce the costs arising from the worker-run problem is to offer select workers different compensation, even though all workers are equally productive. In practice, such differences could be implemented, for instance, by awarding workers somewhat different job titles. Notably, we allow the firm to pay individual team members differently in terms of not only level but also structure. Intuitively, targeting a subset of workers with compensation that makes them willing to stay independent of their coworkers' actions has the advantage of making the remaining workers less concerned about the firm's overall retention. This reduction in strategic complementarities makes it cheaper to retain the remaining workers. Our primary novel insight is that by optimally designing the structure of compensation, the firm can ensure that some workers stay independent of what others do without necessarily paying them more, thereby achieving full retention more cost-effectively.

different job titles

containing contagion effect using rational choice model

To illustrate this insight, suppose that the firm seeks to retain two workers, Alice and Bob, and both workers demand \$100k per year. Suppose further that the firm's overall safe cash flow is only \$100k. Hence, when offering symmetric contracts, the firm must offer some output-dependent pay to both workers. However, since output-dependent pay depends on the firm's success, and thus on its ability to retain both of its productive workers, *both* Alice and Bob will demand compensation for the risk that the other worker leaves. Instead, if the firm combines asymmetric pay with optimal compensation design, the firm does not have to compensate either worker for this risk. The firm can shield the value of Alice's expected compensation from Bob's retention by using its limited resources to guarantee her a payment of $100k via a fixed wage, which is just enough for Alice to always stay. The firm cannot make the same offer to Bob due to its limited resources and, in this stylized example, needs to pay him entirely with output-dependent pay. This compensation structure no longer introduces coordination because Bob no longer has to be concerned that Alice might leave. The more general insight is that when resources are limited, the firm can mitigate the costs arising from the coordination problem, or even avoid them entirely, by structuring workers' compensation differently—paying some workers with a higher share of output-independent pay while offering others more output-dependent pay.[4]

Related Literature. While the substantial cost to firms that arises from worker turnover and the problem of contagious collective turnover are widely discussed by practitioners and the management literature (Felps et al. (2009), Hausknecht and Trevor (2011), Hancock et al. (2013), Heavey, Holwerda, and Hausknecht (2013)), to the best of our knowledge, our paper is the first to formally model this problem and derive implications for firms' compensation policies. Similar to the literature on bank runs (Diamond and Dybvig (1983), Goldstein and Pauzner (2005)), we model inefficient collective turnover as a coordination problem. However, unlike this literature, we focus on compensation design, as standard solutions for bank runs such as deposit insurance, mandatory stay, and suspension of convertibility have no obvious analog in the context of retaining workers.

Our modeling of worker runs as a coordination failure whereby workers leave—even though they would be better off if everyone stayed—highlights that worker runs are inefficient, as workers do not internalize the impact of their decisions on other workers. As illustrated by our example of Credit Suisse, worker runs can also affect reasonably healthy firms and contribute to their collapse: after the Archegos and Greensill Capital disasters, which triggered a worker run in 2021, Credit Suisse's business declined and the bank ultimately suffered a bank run in 2023. This perspective complements work in which workers leave firms because of concerns about their financial health (Titman (1984), Berk, Stanton, and Zechner (2010), Döttling, Ladika, and Perotti (2019)).

A central contribution of our paper is the analysis of dilution—a fundamental feature of equity-based pay that is often neglected in theoretical work.

Is it new?

[4] We show that paying workers differently is less cost-efficient than offering dilutable pay. However, combining asymmetric with dilutable compensation can help if feasibility restrictions constrain the firm's ability to design symmetric dilutable contracts optimally. Moreover, asymmetric compensation is needed by construction if the firm wants to retain some workers with higher probability than others. We discuss optimal compensation design in such settings as an extension.

We show that dilutable equity-based pay can help reduce the sensitivity of workers' compensation to the retention of other workers, which should be particularly important for firms in which keeping teams together is a priority. In line with the evidence, we predict that the dilution feature of equity-based pay is particularly useful in riskier cash-constrained firms, such as startups (Hand (2008)). This helps explain why equity is an effective retention tool despite arguments to the contrary in prior studies that do not account for the dilution benefits of equity pay (Murphy (2003), Lazear (2004), Oyer (2004)).[5] Notably, dilution is not specific to equity-based compensation, but rather is a feature of several types of output-dependent compensation, such as profit-sharing bonus pools at the division or team level or retention bonuses.

The result that firms can reduce the cost of the worker run problem by compensating ex-ante identical workers differently is related to Winter (2004), with whom we share the solution concept of unique implementation. Our central innovation relative to Winter (2004) and subsequent literature is that we explore the idea of offering different agents different types of compensation, such as offering some workers higher fixed pay and others higher performance bonuses. Instead, prior work analyzes the effects of paying some agents more than others while holding the type of compensation fixed. In particular, in Winter (2004) and Halac, Lipnowski, and Rappoport (2021), some agents are offered more call options than others, while in Halac, Kremer, and Winter (2020), some investors are offered a debt-like claim with a higher "face value" than others. In contrast, in our setting in which firms can offer different types of contracts, optimally designed asymmetric compensation does not necessarily introduce substantial differences in pay levels. We further show when contracts that condition on workers' identity (optimal asymmetric nondilutable contract) dominate contracts that condition on the total number of retained workers (optimal symmetric dilutable contracts) in terms of the implied compensation cost. Our analysis of asymmetric contracts further relates to global-games models with heterogeneous payoffs. While early papers in that literature assume that agents' payoff functions are exogenously given (Corsetti et al. (2004), Sakovics and Steiner (2012)), our innovation is to show that firms can benefit from endogenously introducing heterogeneity in payoffs by offering different types of contracts. In recent work, Luo and Yang (2024) have also explored this question by showing that payoff differentiation can emerge endogenously as a designer's solution to the coordination problem in a Sakovics and Steiner (2012) type of global-games setting with optimal security design.

[5] While deferring pay to improve retention is not considered controversial, the question of what type of pay to defer is less clear-cut. Prior theory explains the use of (deferred) equity-based pay with avoiding wage renegotiations when the firm's equity value is correlated with workers' outside options (Oyer (2004)), aligning managers' incentives with investors' interests (Lazear (2004)), exploiting a high stock price in financial markets to attract workers (Bergman and Jenter (2007), Terovitis and Vladimirov (2024)), providing a hedge against not being promoted (Chen (2024)), or hedging Knightian uncertainty (Fulghieri and Dicks (2024)).

I. Model

We develop a tractable compensation design model of a firm seeking to retain a group of skilled workers. All parties are risk-neutral and do not discount future payoffs.

Project. The firm's only asset is a risky project that requires hiring $N \geq 2$ workers indexed by $i \in I = \{1, \ldots, N\}$ at $t = 0$ to get started. Retaining these workers until the end of the project at $t = 2$ is crucial for value creation. In particular, the project generates cash flows at $t = 2$ that can take one of two values: $x > 0$ if the project fails and $x + \Delta x$ with $\Delta x > 0$ if it succeeds. The probability of success $p = p(\varepsilon, n) \in (0, 1)$ depends on the realization of an exogenous shock ε realized at the interim date, $t = 1$, and on the number of workers $0 \leq n \leq N$, the firm can retain following that shock until $t = 2$. The shock ε is drawn from a twice continuously differentiable distribution G with support $[\underline{\varepsilon}, \overline{\varepsilon}]$. We assume that a higher value of ε maps into a higher success probability, $\frac{\partial}{\partial \varepsilon} p(\varepsilon, n) > 0$. Thus, higher realizations of ε reflect positive firm-specific developments or favorable market-wide conditions. For brevity, we refer to ε as a technological shock, but other potential interpretations, such as demand or policy shocks, are also possible. To capture the impact of human capital on project success, we stipulate that p is strictly increasing in n. That is, for all ε and $n < N$, we have that $p(\varepsilon, n) > p(\varepsilon, n-1)$. To simplify exposition, we assume that

$$p(\varepsilon, n) = \alpha(n) + \beta(\varepsilon)\gamma(n), \tag{1}$$

with $\beta(\varepsilon), \beta'(\varepsilon) > 0$ for all ε, $\gamma(n) > 0$ for all n, and $\alpha(n) \in \mathbb{R}$ such that $p(\varepsilon, n) \in [0, 1]$ and $p(\varepsilon, n)$ is increasing in n for all (ε, n). Note that we put few restrictions on the functions α, β, and γ. In particular, our specification is flexible enough for retention and shocks to affect $p(\varepsilon, n)$ both independent of each other as well as through their interaction.[6] This interaction is interesting, as optimal compensation design will differ markedly depending on whether retention and shocks act as complements, $\frac{\partial}{\partial \varepsilon}(p(\varepsilon, n) - p(\varepsilon, n-1)) \geq 0$, or substitutes, $\frac{\partial}{\partial \varepsilon}(p(\varepsilon, n) - p(\varepsilon, n-1)) < 0$.

Interim information and workers' decisions. Workers privately observe the realization of the technological shock ε at $t = 1$. Based on this observation, each worker decides whether to stay with the firm until $t = 2$ or to take an outside option of value $\underline{w}(\varepsilon) \geq 0$. Here, one may think of workers' outside options as the expected payoff from joining another firm or starting a business. Workers make their decisions to stay with or leave the firm simultaneously at $t = 1$ (we discuss sequential decisions in the Internet Appendix).[7] Without loss of generality, we assume that in the case of indifference, a worker prefers to stay with the firm, which could be rationalized with an arbitrarily small switching

[6] For example, suppose that $\gamma(n) = 1 + \widetilde{\gamma}(n)$. The direct effects are then given by $\alpha(n)$ and $\beta(\varepsilon)$, and the interaction is given by $\beta(\varepsilon)\widetilde{\gamma}(n)$. The simplification introduced by (1) is that the dependence of $p(\varepsilon, n)$ on ε is proportional across different values of n, which limits case distinctions.

[7] The Internet Appendix is available in the online version of the article on *The Journal of Finance* website.

cost. Below, it will be useful to distinguish two cases based on how the technological shock ε affects value creation within and outside the firm. We say that the shock ε is *idiosyncratic* if it only affects value creation within the firm, that is, $p(\varepsilon, n)$ is increasing in ε while $\frac{\partial}{\partial \varepsilon}\underline{w}(\varepsilon) = 0$ for all ε. By contrast, we refer to the shock as *systematic* if it also affects workers' outside options, in which case we assume that $\underline{w}(\varepsilon)$ is monotonic in ε, with either $\frac{\partial}{\partial \varepsilon}\underline{w}(\varepsilon) > 0$ or $\frac{\partial}{\partial \varepsilon}\underline{w}(\varepsilon) < 0$ for all ε.

We focus on the problem of *efficient retention* by making the following assumptions. First, we abstract from the problem of talent acquisition. Specifically, we take the number of workers N that the firm hires at $t = 0$ as given and assume that workers do not have a relevant ex-ante participation constraint, that is, their $t = 0$ outside option is not larger than $\mathrm{E}[\underline{w}(\varepsilon)]$. Second, we stipulate that a higher retention level n at $t = 1$ is efficient in the sense of maximizing the expected joint surplus of the firm and its (retained) workers. That is, the expected surplus at $t = 1$,

$$\Omega(\varepsilon, n) := x + p(\varepsilon, n)\Delta x - n\underline{w}(\varepsilon), \tag{2}$$

is positive and nondecreasing in n (i.e., $\Omega(\varepsilon, n) \geq \Omega(\varepsilon, n-1) \geq 0$ for all ε and $1 \leq n \leq N$). The implicit assumption is that the firm cannot replace workers who leave at $t = 1$ with equally productive workers. This assumption captures the problem that finding, hiring, and training replacements for skilled workers is a costly and time-consuming process that lowers the firm's productivity. [8]

Compensation contracts. Compensation contracts are signed at $t = 0$. Workers who leave at $t = 1$ forgo all of their compensation. That is, compensation is deferred, which is always optimal in our setting as it facilitates retention and does not make hiring more difficult. Each worker who stays until $t = 2$ is paid according to compensation contract $C_i := (w_i(n), \Delta w_i(n))_{n=1}^{N}$, which can differ across workers $i \in I$. These contracts stipulate a transfer of $w_i(n)$ to worker i in the low-cash-flow state and of $w_i(n) + \Delta w_i(n)$ in the high-cash-flow state. Going forward, we refer to $w_i(n)$ as output-independent pay and to $\Delta w_i(n)$ as output-dependent pay. Both pay components can condition on the number of workers n that the firm can retain.[9] The key contracting friction is that the technological shock ε is not contractible, as it is only observed privately by workers. Without loss of generality, we abstract from menus of contracts from which workers can choose based on their observation of the interim signal (we discuss menus in the Internet Appendix).

The contracts the firm can offer are subject to the following standard restrictions. Firm owners and workers are protected by limited liability, and contracts are monotone in cash flows. To formally state these restrictions

[8] While not our focus, the problem that a firm might want to reduce the number of workers for efficiency reasons is also important. One way to solve this problem is to make contracts easy to terminate. This solution is common practice in the United States, where most workers are hired at will.

[9] While compensation that explicitly conditions on n may be rare in practice, contracts that do so implicitly are common, including standard time-vesting equity (see Section V for details).

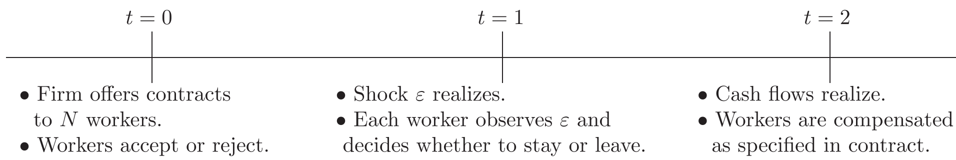

Figure 1. Timeline.

when compensation differs across workers, let $I(n)$ denote the set of all subsets $i(n)$ consisting of n workers. Then, for any n and $i(n) \in I(n)$, limited liability requires that payments be nonnegative and not exceed the firm's available resources $0 \le \sum_{i \in i(n)} w_i(n) \le x$. In turn, the monotonicity constraint, $0 \le \sum_{i \in i(n)} \Delta w_i(n) \le \Delta x$, ensures that both workers and firm owners are at least weakly better off in the high-cash-flow state. The underlying rationale is that no party should have incentives to sabotage the firm by destroying output at $t = 2$ in order to extract higher payoffs (Innes (1990)). We say that contracts C_i for all i are *feasible* if they satisfy limited liability and monotonicity in cash flows, and we denote the set of all feasible contracts by $\mathbf{C}^f$. The firm can commit to any feasible contract at $t = 0$, and all of the above is common knowledge. We discuss the implications of renegotiation possibilities in the Internet Appendix. Figure 1 summarizes our baseline model.

II. Worker Runs—Causes and Potential Remedies

We solve the model backward. In this section, we first take the compensation contract as given and characterize the equilibria of the coordination game at $t = 1$. We then introduce the main contractual features that can help mitigate coordination failure. Below, we explore these features and their usefulness in detail subsequently when solving for optimal contracts as designed by the firm at $t = 0$.

A. Worker Runs as Coordination Failure

Take compensation contract $C_i = (w_i(n), \Delta w_i(n))_{n=1}^{N}$ for each worker i as given and consider workers' decision problem at $t = 1$ after they have observed the shock realization ε. Each worker decides whether to leave or stay with the firm by comparing her expected on-the-job compensation with the available outside option. The key observation underlying our analysis is that this comparison may depend on her beliefs about whether other workers will stay or leave, as the firm's ability to retain workers affects its probability of success.

To build intuition, it is instructive to consider the case in which the firm pays all workers the same symmetric contracts $C_i = C$, for which workers' expected compensation (conditional on retention n), $W(\varepsilon, n) := w(n) + p(\varepsilon, n)\,\Delta w(n)$, increases in the level of retention. One example is a compensation contract that does not condition on the retention level, n, and has a positive output-dependent component, $\Delta w > 0$. Under such a contract, workers' expected compensation net of their outside options, $W(\varepsilon, n) - \underline{w}(\varepsilon)$, increases in the retention level n since the firm's success probability, $p(\varepsilon, n)$, increases in n. Hence, workers play a coordination game with strategic complementarities. An individual worker is more likely to stay if she believes that more of the other workers will stay. As is well known, such games may feature equilibria in which coordination fails. In particular, in our setting, workers might choose to leave because they believe that others are leaving, even though they would prefer to stay if they knew that everyone else was staying. We call a coordination failure of this form a *worker run*.

More precisely, with symmetric contracts, this game has a unique equilibrium only if staying (or respectively, leaving) is a dominant strategy, that is, a worker's preferred choice regardless of the firm's overall retention level. This is the case if either $W(\varepsilon, 1) \geq \underline{w}(\varepsilon)$, such that staying is the dominant strategy, or $W(\varepsilon, N) < \underline{w}(\varepsilon)$, such that leaving is the dominant strategy. In all other cases, that is, whenever $W(\varepsilon, N) \geq \underline{w}(\varepsilon) > W(\varepsilon, 1)$, workers' optimal decisions depend on their beliefs about their coworkers' actions. In such cases, workers do not have dominant strategies, and there exist two symmetric pure-strategy equilibria: a worker-run equilibrium, in which all workers leave the firm and forgo their compensation in favor of their outside option $\underline{w}(\varepsilon)$, and a full-retention equilibrium, in which all workers stay for a payoff of $W(\varepsilon, N)$. Since $W(\varepsilon, N) \geq \underline{w}(\varepsilon)$, all workers are at least weakly worse off in the run equilibrium.

Worker runs also arise if contracts are asymmetric (i.e., differ across workers), but there are two conceptual novelties. First, there are also pure-strategy equilibria in which only subsets of workers run. Second, uniqueness of equilibrium no longer requires that the decision to stay (leave) be dominant for all workers; rather, it need only be *iteratively dominant*, that is, the unique strategy that survives the iterative elimination of strictly dominated strategies. To illustrate what this means, suppose that risk is idiosyncratic (i.e., $\underline{w}(\varepsilon) = \underline{w}$) and there are two workers. Suppose further that both workers are paid according to a contract with $w_i(n) = 0$ and $\Delta w_i(n) = \Delta w_i = \frac{\underline{w}}{p(\underline{\varepsilon}, 2)} > 0$. Under this contract, workers' expected on-the-job pay (conditional on retention n), $W_i(\varepsilon, n) := w_i(n) + p(\varepsilon, n)\Delta w_i(n)$, is higher than their outside options if both workers stay ($W_i(\varepsilon, 2) \geq \underline{w}$), but lower than their outside options if only one worker stays, and the shock realization ε is sufficiently low ($\underline{w} > W_i(\varepsilon, 1)$). For such shock realizations, workers do not have a dominant strategy as they benefit from staying if their coworker stays, but benefit from leaving if their coworker leaves. The lack of a dominant strategy implies that there are two equilibria, one of which is a worker run. The key point is that if only worker $i = 2$ is paid according to this contract, while worker $i = 1$ receives a fixed-wage contract that matches her outside option exactly, that is, $w_1(n) = \underline{w}$ and $\Delta w_1(n) = 0$ for all n, the unique equilibrium is that both workers stay. Hence, the worker-run equilibrium is no longer supported. Although staying is still not a dominant strategy for worker $i = 2$, it is an *iteratively* dominant strategy because she can condition her action on the fact that worker $i = 1$ has a dominant strategy to stay. Summarizing, we have the following result.

PROPOSITION 1: *Take a feasible collection of compensation contracts* $\{C_i\}_{i=1}^N \in \mathbf{C}^f$ *as given. Denote the number of workers who have an iteratively dominant strategy to stay or leave under this set of contracts by* N^s *and* N^l*, respectively. Then workers' coordination game at* $t = 1$ *has a unique equilibrium if and only if* $N^l + N^s = N$*. Otherwise, there is a subset of workers for which*

$$W_i(\varepsilon, N - N^l) \geq \underline{w}(\varepsilon) > W_i(\varepsilon, N^s + 1), \tag{3}$$

whose optimal decisions to stay or leave depend on their beliefs about overall retention. In this case, there exist a multiplicity of equilibria that involve at least one worker-run equilibrium in which strictly more than N^l *workers leave. Any worker-run equilibrium is Pareto-dominated by the equilibrium in which* N^l *workers leave and* $N - N^l$ *workers stay.*

B. Contractual Features Mitigating the Risk of Worker Runs

Proposition 1 identifies condition (3) as necessary and sufficient for a worker-run equilibrium to exist for a given shock realization ε. According to this condition, two basic compensation features expose the firm to potential runs: (i) expected compensation depends on the level of retention and (ii) expected compensation is neither too high nor too low. In particular, the firm can always retain workers without the risk of a run by sufficiently raising the level of pay. Alternatively, the firm can also avoid the risk of a worker run by designing the structure of compensation such that expected on-the-job compensation $W_i(\varepsilon, n)$ is insensitive to retention n, eliminating the strategic complementarities in workers' decisions to leave. The benefit of structuring compensation in this way is that it allows the firm to reduce the expected level of pay needed to achieve retention. Before analyzing optimal compensation design in detail, we preview the three main compensation structures that help the firm make expected compensation less sensitive to retention.

1. *Fixed pay*. The most straightforward way to make expected compensation independent of the level of retention is via a riskless *fixed-wage* contract that has an output-dependent component of zero ($\Delta w_i(n) = 0$ for all i, n) and an output-independent component that does not depend on the level of retention ($w_i(n) = w_i$ for all i, n). In particular, if $w_i \geq \underline{w}(\varepsilon)$, workers have a dominant strategy to stay for the given shock realization ε.
2. *Dilutable pay*. Contracts involving output-dependent pay expose workers to the firm's success probability and thus the firm's retention level. Yet, such contracts can also be structured to make a worker's expected on-the-job compensation less sensitive to—or even independent of—overall retention n.[10] The key is to make compensation "*dilutable*," that is, to promise a worker who stays less when (overall) retention is higher and more when retention is lower.

[10] In our setting, firms offer output-dependent pay because either the firm's resource constraint is binding or they seek to extract rent.

DEFINITION 1: Compensation of worker i is dilutable at (ε, n) if, holding the probability of success constant at any given $\widehat{p} \in [p(\varepsilon, n-1), p(\varepsilon, n)]$, expected compensation $\widehat{W}_i(\widehat{p}, n) := w_i(n) + \widehat{p}\Delta w_i(n)$ decreases in retention:

$$\widehat{W}_i(\widehat{p}, n) - \widehat{W}_i(\widehat{p}, n-1) = w_i(n) - w_i(n-1) + \widehat{p}[\Delta w_i(n) - \Delta w_i(n-1)] < 0. \quad (4)$$

We refer to the strength of this decrease scaled by the compensation level at n, $\frac{\left|\widehat{W}(\widehat{p},n)-\widehat{W}(\widehat{p},n-1)\right|}{\widehat{W}(\widehat{p},n)}$, as the degree of dilution.

Dilution is a feature of various compensation practices that are common in practice. As an illustration, consider the example from the introduction in which a firm with two workers and 100 shares outstanding sets aside 100 additional shares for compensation purposes, promising each worker 50 shares conditional on staying until $t = 2$. If one worker leaves prematurely, that is, before her shares vest at $t = 2$, the remaining worker's 50 shares amount to a fraction of $\frac{50}{150} = 33\%$ of the firm's outstanding equity, which is larger than the $\frac{50}{200} = 25\%$ equity stake she would hold under full retention. Intuitively, while the firm's total equity value—the "size of the pie"—changes in the firm's retention level, the workers' percentage of equity ownership—their "share of the pie"—adjusts so as to smooth workers' expected compensation across different retention states. In particular, workers' equity share is diluted in high-retention scenarios, in which the value of equity is high and the firm wants to avoid overpaying workers, while it increases in low-retention scenarios, in which the value of equity is low and the firm wants to avoid further worker departures. As previously noted, dilution is not specific to equity-based compensation. For example, profit-sharing bonus pools at the division or team level (tying dilution to team size) and retention bonuses are other simple ways to make compensation dilutable.[11]

More generally, a necessary condition for compensation to be dilutable is that either $w_i(n)$ or $\Delta w_i(n)$ decreases in the level of retention n; a sufficient condition is that both $w_i(n)$ and $\Delta w_i(n)$ decrease in n, as in the equity example above. Regardless of how it is achieved, dilution is a necessary countervailing force to ensure that workers' expected on-the-job compensation does not decrease when the firm's success probability drops as overall retention decreases.

LEMMA 1: *Any contract with $\Delta w_i(n) > 0$ such that worker i's expected compensation does not decrease at ε as n decreases (i.e., $W_i(\varepsilon, n-1) \geq W_i(\varepsilon, n)$) must be dilutable at (ε, n).*

3. *Asymmetric pay*. An alternative way to eliminate strategic complementarities in the worker retention problem is to offer asymmetric contracts that differ across workers, where our focus is on differences in compensation *structure*. We have already illustrated the benefit of asymmetric

[11] To implement more complex dilution patterns, the firm may combine time- and performance-vesting provisions, offer retention bonuses, or rely on equity buy-back agreements (see Section V for details).

compensation in our discussion of iteratively dominant strategies preceding Proposition 1. In this two-worker example, offering worker 1 a fixed-wage contract equal to her outside option fully eliminates strategic complementarities. The reason is that worker 1 then has a dominant strategy to stay, such that the only relevant retention level for worker 2 is $n = 2$. Hence, paying worker 2 with $(w_2, \Delta w_2) = \left(0, \frac{w}{p(\underline{\varepsilon},2)}\right)$ avoids worker runs even though this same contract would have resulted in worker-run equilibria if offered to both workers.

The discussion so far illustrates that the firm can employ a variety of compensation design tools available to tackle the coordination problem among workers. Next, we analyze how the firm optimally makes use of this rich toolset. Specifically, proceeding backward to $t = 0$, we next characterize optimal *full-retention* contracts that avoid worker runs at $t = 1$ for all realizations of the technological shock ε at the lowest compensation cost.[12] In Section III, we focus on symmetric contracts, while Section IV considers asymmetric contracts.

III. Symmetric Fixed and Dilutable Contracts

We start our analysis of optimal full-retention contracts by restricting attention to symmetric contracts, that is, $C_i = C$ for all i. As we will show, doing so is without loss of generality if feasibility constraints are slack, and it allows us to highlight the use of fixed and especially dilutable compensation in optimal contracts in the clearest possible way. We characterize optimal symmetric contracts—establishing dilution as an optimal contracting feature—in Section III.A. In Section III.B, we analyze the different dilution patterns that can arise. Sections III.C and III.D discuss further features of optimal symmetric compensation design and the role of feasibility constraints in shaping optimal contracts.

A. Optimality of Fixed and Dilutable Compensation

In our model, the objective of optimal compensation design at $t = 0$ is to minimize workers' expected rent—defined as their expected compensation net of outside options—while guaranteeing full retention. To implement full retention as the unique Nash equilibrium, a *symmetric* contract must ensure that all individual workers are better off staying for all possible shock realizations ε and retention levels n. That is, the firm minimizes workers' equilibrium expected rent subject to the constraints that their interim ($t = 1$) expected compensation must be larger than their outside option both on-equilibrium

[12] Solving the firm's problem for retention policies other than full retention requires an extension of our model to address the multiplicity of equilibria described in Proposition 1. One such extension, which we discuss in Section IV.C and formally analyze in the Internet Appendix, is to impose a standard global games refinement in which workers observe ε only with noise.

(for $n = N$) and off-equilibrium (for $n < N$) to ensure that staying is a dominant strategy in the presence of coordination frictions (Proposition 1). To simplify notation, we drop the subscript i in this section since workers and contracts are symmetric.[13]

Problem 1 (Optimal symmetric full retention contract)**:**

$$\min_{C \in \mathbf{C}^f} \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \left[w(N) + p(\varepsilon, N)\Delta w(N) - \underline{w}(\varepsilon)\right] dG(\varepsilon) \tag{5}$$

$$s.t.\ W(\varepsilon, n) = w(n) + p(\varepsilon, n)\Delta w(n) \geq \underline{w}(\varepsilon)\ \forall(\varepsilon, n). \tag{6}$$

The benchmark against which we can measure the cost stemming from the coordination friction in Problem 1 is the case in which workers act as one in the sense of deciding whether to stay or leave jointly as a group. In this "*relaxed problem*," the firm continues to minimize compensation costs in (5) but only subject to the full-retention participation constraints, $W(\varepsilon, N) \geq \underline{w}(\varepsilon)$, since compensation for lower retention levels $n < N$ is irrelevant absent coordination frictions. Denote the solution to this relaxed problem by $(w^*(N), \Delta w^*(N))$ and each worker's associated expected compensation at ε by $W^*(\varepsilon, N) := w^*(N) + p(\varepsilon, N)\Delta w^*(N)$. The existence of such a solution follows standard arguments. We refer to this solution as "interior" whenever feasibility constraints are slack in the relaxed problem.[14]

For now, we take the benchmark solution to the relaxed problem as given since our focus is on the optimal contractual solution to the coordination problem, the cost of which can be measured by how close the firm can get to this benchmark. In particular, if the firm can achieve the same compensation costs in the full problem with coordination frictions (Problem 1) as in the relaxed problem, then the firm can resolve the coordination problem at no cost. Clearly, this requires setting compensation under full retention exactly as in the relaxed problem, $\big(w(N), \Delta w(N)\big) = \big(w^*(N), \Delta w^*(N)\big)$. We now argue that this is always possible, but requires either fixed or dilutable pay to ensure that $\big(w(n), \Delta w(n)\big)$ for $n < N$ satisfies the remaining off-equilibrium participation constraints.

PROPOSITION 2: *Suppose that the firm wants to implement full retention as the unique equilibrium using a symmetric contract, and let* $\big(w^*(N), \Delta w^*(N)\big)$ *denote the optimal contract in the relaxed problem in which workers can coordinate. Then, the firm can achieve the same expected compensation costs in the full problem, that is, can resolve the coordination problem at no additional cost. The optimal contract stipulates compensation conditional on full retention of* $\big(w(N), \Delta w(N)\big) = \big(w^*(N), \Delta w^*(N)\big)$ *and sets compensation for* $n < N$ *to satisfy the following features:*

[13] In a slight abuse of notation, we thus write $C \in \mathbf{C}^f$ if $\{C_i\}_{i=1}^N \in \mathbf{C}^f$ and $C_i = C$ for all i.

[14] Formally, this means that the constraints $0 \leq Nw(N) \leq x$ and $0 \leq N\Delta w(N) \leq \Delta x$ have shadow costs of zero in the relaxed problem. We discuss the determination of $(w^*(N), \Delta w^*(N))$ in Section III.C.

(i) If $\Delta w^*(N) > 0$*, any optimal contract must be dilutable at N on a set of* ε *with positive measure.*

(ii) An optimal contract that is not dilutable at any (ε, n) *exists if and only if* $\Delta w^*(N) = 0$*; in this case, a fixed wage with* $w(n) = w^*(N)$ *and* $\Delta w(n) = \Delta w^*(N) = 0$ *for all n constitutes an optimal contract.*

An optimal contract that is dilutable at any (ε, n) *always exists.*

The dilution properties described in statements (i) and (ii) of Proposition 2 follow from a simple observation. Absent coordination frictions, the optimal full-retention contract $(w^*(N), \Delta w^*(N))$ will satisfy workers' participation constraint with equality for at least one shock realization. This means that if the firm offers workers the same contract when there are coordination frictions, the workers will be worried about whether the firm can manage to maintain full retention (and thus a high probability of success) for that shock realization. To alleviate these concerns, the firm must promise workers an adjustment to their compensation if retention drops to $n < N$ to ensure that staying remains preferable. That is, the firm must offer a dilutable contract (Lemma 1). The only exception is if the firm optimally pays workers purely with fixed compensation.

Proposition 2 further argues that a dilutable contract resolving the coordination problem at no cost at all (ε, n) always exists. A trivial example is a contract that promises all cash flows to workers, that is, $(w(n), \Delta w(n)) = \left(\frac{x}{n}, \frac{\Delta x}{n}\right)$ for out-of-equilibrium retention levels $n < N$. This contract is dilutable since both $w(n)$ and $\Delta w(n)$ decrease in n for all n (note that $w^*(N) \leq \frac{x}{N}$ and $\Delta w^*(N) \leq \frac{\Delta x}{N}$). Furthermore, since the firm generates a surplus over workers' outside options, $\Omega(\varepsilon, n) \geq 0$ for all (ε, n), it is a dominant strategy for workers to stay under this contract. The contract also trivially satisfies the limited liability and monotonicity-in-cash-flow constraints, and hence is always feasible.

While the contract that we have just described avoids worker runs at no cost, it has the intuitively undesirable feature that workers may benefit financially when their coworkers depart during the project's implementation phase at $t = 1$. In richer settings that capture realistic workplace dynamics, such contracts are likely problematic because they can contribute to a toxic work environment, as individual workers stand to benefit from inducing their coworkers to leave, for example, by not collaborating properly or even resorting to mobbing. Such contracts therefore appear unsuitable for promoting effective teamwork, which is vital for firm success in several of our key applications. To rule out these unrealistic contracts, we impose a contracting restriction that requires expected compensation to monotonically increase in retention:[15]

[15] This restriction is similar in spirit to standard financial contracting restrictions. In particular, Innes (1990) motivates the restriction that contractual payoffs must be monotone in cash flows with an ex post moral hazard problem that agents should have no incentives to destroy cash flows *after* cash flows are realized at $t = 2$. By contrast, we are interested in workers' interim expected compensation, as workers make the decision to leave, and the firm tries to retain workers during the project's implementation phase at $t = 1$ *before* cash flows are realized. A separate motivation

ASSUMPTION 1 (Monotonicity in retention): *Contracts must be monotonic in retention; that is, a worker's expected on-the-job pay* $W_i(\varepsilon, n) := w_i(n) + p(\varepsilon, n)\,\Delta w_i(n)$ *must be nondecreasing in n (i.e.,* $W_i(\varepsilon, n) \geq W_i(\varepsilon, n-1)$*), for all* (ε, n) *and i.*

Monotonicity in retention requires that workers should not benefit when other workers leave during the project's implementation phase, $t = 1$. In other words, monotonicity in retention puts an upper bound on the degree of dilution by restricting the extent to which promised compensation can increase when retention drops. Intuitively, optimal contracts satisfying Assumption 1 can still be dilutable, but contracts "overdosing" on dilution are ruled out. In that sense, the monotonicity-in-retention constraint effectively acts as a refinement.

To see this, consider the implications of augmenting Problem 1 with the monotonicity-in-retention constraint. Together with the interim participation constraints (6), the monotonicity-in-retention constraint implies that all admissible contracts must satisfy

$$W(\varepsilon, N) \geq \ldots \geq W(\varepsilon, 1) \geq \underline{w}(\varepsilon)\ \forall (\varepsilon, n). \tag{7}$$

It is now convenient to write workers' equilibrium rent, $R(\varepsilon, N) := W(\varepsilon, N) - \underline{w}(\varepsilon)$, as the sum of the "incremental rents" at each $n \geq 1$,

$$R(\varepsilon, N) = \left[W(\varepsilon, N) - W(\varepsilon, N-1)\right] + \cdots + \left[W(\varepsilon, 2) - W(\varepsilon, 1)\right] + \left[W(\varepsilon, 1) - \underline{w}(\varepsilon)\right] \geq 0, \tag{8}$$

where each incremental rent (in square brackets) must be nonnegative by condition (7). Hence, the objective of minimizing workers' expected rent $\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} R(\varepsilon, N) dG(\varepsilon)$ subject to (7) pushes toward a contract that sets as many incremental rents to zero as possible. It therefore becomes optimal to hold expected compensation constant in retention, that is, $W(\varepsilon, n-1) = W(\varepsilon, n)$ for all (ε, n). Intuitively, doing so perfectly smooths workers' expected compensation across different retention scenarios and shock realizations. This requires choosing the output-dependent components $\Delta w(n)$ such that the sensitivity of workers' expected compensation to shocks, $\frac{\partial W(\varepsilon,n)}{\partial \varepsilon} = \frac{\partial p(\varepsilon,n)}{\partial \varepsilon}\Delta w(n)$, is constant in retention or, equivalently, $\Delta w(n-1) = \frac{\partial p(\varepsilon,n)/\partial \varepsilon}{\partial p(\varepsilon,n-1)/\partial \varepsilon}\Delta w(n)$ must hold for all retention levels n. Substituting for the functional form of $p(\varepsilon, n)$ from (1) and determining the output-independent component as a residual from $W(\varepsilon, n-1) = W(\varepsilon, n)$, we have that

$$\Delta w(n-1) = \frac{\gamma(n)}{\gamma(n-1)}\Delta w(n), \tag{9}$$

$$w(n-1) = w(n) + \left[\alpha(n) - \alpha(n-1)\frac{\gamma(n)}{\gamma(n-1)}\right]\Delta w(n). \tag{10}$$

is that the firm would clearly eschew contracts violating Assumption 1 if there were a probability that workers would leave for exogenous reasons, such as health or family reasons.

If the iterative construction prescribed by (9) and (10) is feasible when choosing compensation under full retention to be the same as in the relaxed problem (i.e., $\left(w^*(N), \Delta w^*(N)\right)$), the monotonicity-in-retention constraint has zero shadow cost and the coordination problem entails no cost.

PROPOSITION 3: *A symmetric contract C^*, which pays each worker $\left(w^*(N), \Delta w^*(N)\right)$ conditional on full retention, $n = N$, and specifies payments of $\left(w^*(n), \Delta w^*(n)\right)$ for all lower retention levels $n < N$ according to (9) and (10), implements full retention as the unique equilibrium at no coordination costs. This contract holds expected compensation $W(\varepsilon, n)$ constant in n for each ε—and thus is monotonic in retention—which is achieved by offering dilutable compensation for all (ε, n) if $\Delta w^*(N) > 0$ (dilutable-pay contract) or by paying only a fixed wage if $\Delta w^*(N) = 0$ (fixed-wage contract). Hence, if feasible (i.e., $0 \le w^*(n) \le \frac{x}{n}$ and $0 \le \Delta w^*(n) \le \frac{\Delta x}{n}$ for all n), contract C^* solves the optimal compensation design Problem 1 subject to the monotonicity-in-retention refinement, uniquely so if $\left(w^*(N), \Delta w^*(N)\right)$ is interior.*

Proposition 3 shows both the optimality of smoothing workers' expected compensation across different retention levels, $W(\varepsilon, n) = W(\varepsilon, N)\ \forall (\varepsilon, n)$, and how the optimal contract achieves that smoothing, which strengthens the results obtained without the monotonicity-in-retention constraint in Proposition 2. In particular, if the firm prefers to pay workers a fixed wage conditional on full retention, then smoothing workers' compensation requires paying workers the same fixed wage for all other retention levels $n < N$.

If instead workers optimally receive positive output-dependent compensation under full retention, $\Delta w^*(N) > 0$, then holding workers' expected compensation constant in n for all ε requires offering a contract with a positive output-dependent component that is dilutable everywhere. This claim is easy to see from (9), which implies that an increase in $\Delta w^*(N)$ results in an increase in $\Delta w(n)$ for all $n < N$. Below, we discuss in more detail the closed-form solutions provided by the iterative construction in (9) and (10) for a given $(w^*(N), \Delta w^*(N))$, as they allow us to analyze the determinants of the optimal degree of dilution and different dilution patterns of the contract from Proposition 3.

B. Degree of Dilution and Dilution Patterns

The degree of dilution needed to perfectly smooth expected compensation across retention states, as prescribed by Proposition 3, is determined by two main factors. First, workers' expected compensation is more exposed to success in production, and therefore requires a higher degree of dilution if the share of output-dependent to total pay $\Delta w^*(n)/\left(w^*(n) + \Delta w^*(n)\right)$ is higher. As the preceding discussion indicates, this share increases for all retention levels in the corresponding share under full retention, $\Delta w^*(N)/\left(w^*(N) + \Delta w^*(N)\right)$, as pinned down by the solution of the relaxed problem. Second, since dilution seeks to balance the effect of retention on workers' expected compensation via the production technology, dilution is needed more as a countervailing force if the firm's probability of success, $p(\varepsilon, n)$, is more sensitive to retention. [16]

COROLLARY 1: *Fix any* $\varepsilon \in [\underline{\varepsilon}, \overline{\varepsilon}]$ *and a* $\widehat{p} \in (p(\varepsilon, n-1), p(\varepsilon, n))$ *for each* $n \geq 2$*, and consider the optimal contract characterized in Proposition 3 with* $\Delta w^*(N) > 0$*. The degree of dilution,* $\frac{|\widehat{W}(\widehat{p},n)-\widehat{W}(\widehat{p},n-1)|}{\widehat{W}(\widehat{p},n)}$*, at all* $n \geq 2$ *is strictly increasing in the equilibrium share of output-dependent pay under full retention,* $\frac{\Delta w^*(N)}{w^*(N)+\Delta w^*(N)}$*. Furthermore, the degree of dilution at any given* $n \geq 2$ *increases if* $\alpha(n-1)$ *or* $\gamma(n-1)$ *decrease, such that from (1)* $p(\varepsilon, n-1)$ *decreases relative to the fixed* $p(\varepsilon, n)$*.*

Different Dilution Patterns. We now explore the different dilution patterns predicted by Proposition 3. In particular, we explore how the compensation components, $w^*(n)$ and $\Delta w^*(n)$, change in the firm's retention level, n, to make workers' expected compensation, $W(\varepsilon, n)$, insensitive to retention n. Note from $\Delta w^*(n) = \frac{\partial p(\varepsilon,n-1)/\partial\varepsilon}{\partial p(\varepsilon,n)/\partial\varepsilon}\Delta w^*(n-1)$ (which corresponds to (9)) that whether output-dependent pay, $\Delta w^*(n)$, increases or decreases in retention depends on whether shocks have a stronger impact on the firm's probability of success at high retention levels, ($\frac{\partial p(\varepsilon,n-1)/\partial\varepsilon}{\partial p(\varepsilon,n)/\partial\varepsilon} = \frac{\gamma(n-1)}{\gamma(n)} < 1$) or low retention levels ($\frac{\partial p(\varepsilon,n-1)/\partial\varepsilon}{\partial p(\varepsilon,n)/\partial\varepsilon} \geq 1$). This distinction corresponds to whether technological shocks and retention are complements ($\frac{\partial}{\partial\varepsilon}\left(p(\varepsilon, n) - p(\varepsilon, n-1)\right) > 0$) or substitutes ($\frac{\partial}{\partial\varepsilon}\left(p(\varepsilon, n) - p(\varepsilon, n-1)\right) \leq 0$).

In the case of *complements*, shocks have a higher impact on the firm's probability of success at higher retention levels. Thus, to keep workers' exposure to shocks, $\frac{\partial W(\varepsilon,n)}{\partial\varepsilon} = \frac{\partial p(\varepsilon,n)}{\partial\varepsilon}\Delta w(n)$, constant in retention n, optimal output-dependent compensation, $\Delta w^*(n)$, must be decreasing in n. Whether the output-independent component $w^*(n)$ also decreases in n then depends on whether the dilution achieved through $\Delta w^*(n)$ is sufficient to smooth $W(\varepsilon, n)$ across n. If the degree of dilution achieved is insufficient, $w^*(n)$ will also decrease in n (Panel A.3 in Figure 2). If it is just sufficient, $w(n)$ will be flat in n (Panel A.2) while if it is excessive, $w^*(n)$ will increase in n (Panel A.1).

In the case of *substitutes*, shocks have a lower impact on the firm's probability of success at higher retention levels. Thus, to keep workers' exposure to shocks constant in n, the firm must offer output-dependent compensation, $\Delta w^*(n)$, that increases in n. To make the overall contract dilutable, the firm must therefore make output-independent pay $w^*(n)$ decreasing in n (see Panels B and C in Figure 2).[17]

COROLLARY 2: *Consider the optimal contract* C^* *from Proposition 3 with* $\Delta w^*(N) > 0$*.*

[16] While Corollary 1 establishes a local version of this result, one can also obtain a global analog. To see this, assume we can write $\alpha(n-1) = \frac{1}{\kappa_\alpha}\alpha(n)$ and $\gamma(n-1) = \frac{1}{\kappa_\gamma}\gamma(n)$ for all $n \geq 2$ and some $\kappa_\alpha > 1$ and $\kappa_\gamma > 1$. Then, if κ_α or κ_γ increases, that is, from (1) the probability of success becomes more sensitive to retention everywhere, the degree of dilution increases for all $(\varepsilon, n \geq 2)$. For details, see the proof of Corollary 1.

[17] Panel B in Figure 2 corresponds to the case in which $\frac{\partial p(\varepsilon,N)/\partial\varepsilon}{\partial p(\varepsilon,n)/\partial\varepsilon} = 1$, and so, $\Delta w(n)$ is flat in n.

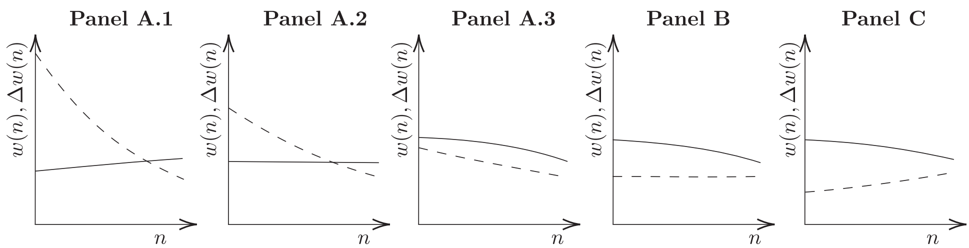

Figure 2. Dilutable contracts under full retention. The figure plots different dilution patterns for contracts satisfying $W(\varepsilon, n) = W(\varepsilon, N)$ for all (ε, n). The solid curve represents $w(n)$ and the dashed curve $\Delta w(n)$. Panels A.1 to A.3 depict cases in which technological shocks have a stronger impact at higher retention levels ($\partial p(\varepsilon, n)/\partial\varepsilon$ increasing in n), differentiating between a strong, moderate, and weak impact of the shocks. In Panel B, shocks have the same impact across different retention levels ($\partial p(\varepsilon, n)/\partial\varepsilon$ independent of n). Panel C depicts the case in which shocks have a stronger impact at lower retention levels ($\partial p(\varepsilon, n)/\partial\varepsilon$ decreasing in n).

(i) Complements: If the impact of technological shocks on the firm's success probability is increasing in the firm's retention level, $\frac{\partial}{\partial\varepsilon}\big(p(\varepsilon, n) - p(\varepsilon, n-1)\big) > 0$, dilution is achieved by making output-dependent pay $\Delta w^*(n)$ decreasing in n. Output-independent pay $w^*(n)$ is decreasing in n, and thus contributes to dilution, if $\frac{\gamma(n)}{\gamma(n-1)}$ is below a cutoff; it is increasing in n otherwise.

(ii) Substitutes: If the impact of technological shocks on the firm's success probability is decreasing in the firm's retention level, $\frac{\partial}{\partial\varepsilon}\big(p(\varepsilon, n+1) - p(\varepsilon, n)\big) < 0$, output-dependent pay, $\Delta w^*(n)$, is increasing in n. To make overall compensation dilutable, the firm makes output-independent pay, $w^*(n)$, sufficiently decreasing in n.

Corollary 2 and its graphical illustration in Figure 2 foreshadow the rich empirical implications of dilution for compensation design. In particular, in Section V, we discuss how the firm can implement these patterns through time-vesting equity-based pay (Panel A.3 in Figure 2), performance-vesting equity-based pay or bonus pools (Panel A.2), equity buyback agreements (Panel A.1), and retention bonuses (Panels B and C).

C. Output-Independent and Output-Dependent Pay Under Full Retention

So far, our compensation-design analysis has taken as given the solution to the relaxed problem in which workers can coordinate, $\big(w^*(N), \Delta w^*(N)\big)$, and shown how an optimally designed dilutable or fixed-wage contact can achieve the same compensation costs in the full problem with coordination frictions. We now close the analysis of optimal compensation design by studying how $\big(w^*(N), \Delta w^*(N)\big)$, that is, the optimal mix of output-dependent and output-independent pay conditional on full retention, is determined. To do so, we initially abstract from relevant feasibility constraints, which we turn to in the next section. In either case, the optimal mix of output-dependent and output-independent pay is driven by the firm's objective of minimizing the rent workers extract due to the fact that the technological shock ε is noncontractible.[18]

Recall that the firm's objective in the relaxed problem in which workers can coordinate is to choose a feasible contract that minimizes workers' expected rent $\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \left(W(\varepsilon, N) - \underline{w}(\varepsilon)\right) dG(\varepsilon)$ subject only to the full-retention participation constraint $W(\varepsilon, N) \geq \underline{w}(\varepsilon)$ for each ε. The solution to this problem is particularly simple if risk is idiosyncratic, $\underline{w}(\varepsilon) = \underline{w}$ for all ε, and the available resources in the low-cash-flow state are sufficiently large, $x \geq N\underline{w}$. In this case, offering only output-independent pay, $w^*(N) = \underline{w}$, and no output-dependent pay, $\Delta w^*(N) = 0$, extracts all rent from workers. If instead risk is systematic with $\frac{\partial \underline{w}(\varepsilon)}{\partial \varepsilon} > 0$, the firm's rent extraction motive calls for on-the-job compensation that increases after positive shocks—in which case workers' outside options are high—and decreases following negative shocks—in which case workers' outside options are low. That is, the optimal contract calls for at least some output-dependent compensation, $\Delta w^*(N) > 0$.[19] More generally, the objective of matching workers' expected on-the-job compensation as closely as possible to their outside options implies that the share of output-dependent compensation, $\Delta w^*(N)/\left(w^*(N) + \Delta w^*(N)\right)$, should increase as the sensitivity of workers' outside options to shocks, $\frac{\partial}{\partial \varepsilon}\underline{w}(\varepsilon)$, increases relative to the sensitivity of firm output to shocks, $\frac{\partial}{\partial \varepsilon}p(\varepsilon, n)$. Continuing to abstract for now from feasibility issues (i.e., considering an "interior solution"), we have the following result.

PROPOSITION 4: *At an interior solution, the optimal share of output-dependent pay in workers' compensation conditional on full retention,* $\frac{\Delta w^*(N)}{w^*(N)+\Delta w^*(N)}$*, is determined by the sensitivity of workers' outside option to technological shocks,* $\frac{\partial}{\partial \varepsilon}\underline{w}(\varepsilon)$*, relative to the sensitivity of the firm's production technology to such shocks,* $\frac{\partial}{\partial \varepsilon}p(\varepsilon, N)$*. Specifically, denote by* $\epsilon_N \subseteq \left[\underline{\varepsilon}, \overline{\varepsilon}\right]$ *the set of points at which workers' participation constraint holds with equality,* $W^*(\varepsilon, N) = \underline{w}(\varepsilon)$*. If, at some point* $\varepsilon_N \in \epsilon_N$*,* $W^*(\varepsilon, N)$ *and* $\underline{w}(\varepsilon)$ *are tangent to each other, then* $\Delta w^*(N) = \frac{\partial}{\partial \varepsilon}\underline{w}(\varepsilon) / \frac{\partial}{\partial \varepsilon}p(\varepsilon, N)\Big|_{\varepsilon=\varepsilon_N}$*. Otherwise,* $\Delta w^*(N) = \left(\underline{w}(\overline{\varepsilon}) - \underline{w}(\underline{\varepsilon})\right) / \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \frac{\partial}{\partial \varepsilon}p(\varepsilon, n)d\varepsilon$*. Finally,* $w^*(N) = \underline{w}(\varepsilon_N) - p(\varepsilon_N, N)\Delta w^*$ *for any* $\varepsilon_N \in \epsilon_N$*.*

[18] If shocks were contractible, there would be infinitely many contracts $(w(\varepsilon, n), \Delta w(\varepsilon, n))_{n=1}^N$ (including equity, call options, and fixed wages) for which workers' on-the-job pay is equal to their outside option, $W(\varepsilon, n) = \underline{w}(\varepsilon)$, for all (ε, n). In this case, the only prediction about optimal contract design is that all such contracts (which are not fixed wages) are dilutable at all (ε, n) (see Lemma 1).

[19] If $\frac{\partial}{\partial \varepsilon}\underline{w}(\varepsilon) < 0$, shocks that are good for the firm are bad for workers' outside options. In this case, the firm would like to make $W(\varepsilon, n)$ decreasing in ε to minimize workers' rent. However, doing so is infeasible since $\frac{\partial}{\partial \varepsilon}W(\varepsilon, n) = \frac{\partial}{\partial \varepsilon}p(\varepsilon, n)\Delta w(n) < 0$ requires $\Delta w(n) < 0$, violating the monotonicity-in-cash-flows constraint $\Delta w(n) \geq 0$. Hence, the optimal contract is not "interior," but rather is determined by the binding constraint, and is again given by a fixed-wage with $w(n) = \underline{w}(\underline{\varepsilon})$ and $\Delta w(n) = 0$ for all n, which is feasible as long as $N\underline{w}(\underline{\varepsilon}) < x$.

D. Binding Feasibility Constraints

In Propositions 3 and 4 above, we study the case in which feasibility constraints are slack. To study the impact of binding feasibility constraints, recall that the firm's objective in Problem 1 is to choose a feasible contract that minimizes workers' expected equilibrium rent $\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} R(\varepsilon, N)\, dG(\varepsilon)$, subject to the restrictions imposed by their interim participation and monotonicity-in-retention constraints as captured by (7). Hence, at any retention level n, the firm will choose payments $\big(w(n), \Delta w(n)\big)$ to minimize the (downward) incremental rent $W(\varepsilon, n) - W(\varepsilon, n-1)$ in (8). If it is feasible to push this rent to zero for all ε, this is clearly optimal. In this case, the relation between $\Delta w\,(n)$ and $\Delta w\,(n-1)$ is again given by (9) and that between $w\,(n)$ and $w\,(n-1)$ by (10) as in the zero-incremental-rent contract of Proposition 3, which leads to the same dilution properties. If it is not feasible to push the incremental rent to zero for all ε, feasibility constraints codetermine the shape of the optimal contract.[20] In this case, the dilution predictions from Corollaries 1 and 2 remain valid as long as feasibility restrictions are not too strict. Notably, setting the incremental rent $W(\varepsilon, n) - W(\varepsilon, n-1)$ to zero for at least some ε is always feasible at any level of retention.[21] By Lemma 1, it follows that fixed and dilutable compensation also remain a robust feature of optimal contracts at any retention level when feasibility constraints bind. We provide a formal characterization and an algorithm for constructing the optimal contract when feasibility constraints bind in the Internet Appendix.

The key economic difference if feasibility constraints are sufficiently binding is that optimal symmetric contracting may no longer be able to resolve the coordination problem at no cost. As is clear from expression (8), this case of costly coordination occurs if and only if the sum of expected incremental rents is strictly larger than workers' expected rent, $\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \big(W^*(\varepsilon, N) - \underline{w}(\varepsilon)\big)\, dG(\varepsilon)$, in the relaxed problem in which there is no coordination problem. Given that resolving the coordination problem at no cost is always possible if the monotonicity-in-retention constraint is not imposed or if feasibility constraints are slack (Propositions 2 and 3), we have the following lemma.

LEMMA 2: *Optimal symmetric contracts entail strictly positive coordination costs if and only if feasibility constraints cause the monotonicity-in-retention constraint to bind.*

If binding feasibility constraints cause coordination costs, the firm can potentially do better by offering asymmetric contracts.

[20] For example, if expression (9) violates cash-flow monotonicity, $\Delta w\,(n) \leq \frac{\Delta x}{n}$, the firm optimally sets $\Delta w\,(n) = \frac{\Delta x}{n}$. Similarly, if expression (10) violates limited liability, $w\,(n) \geq 0$, the firm offers $w\,(n) = 0$, and if expression (10) violates the resource constraint $w\,(n) \leq \frac{x}{n}$, the firm will offer $w\,(n) = \frac{x}{n}$.

[21] Suppose that this was not the case. Then, one could reduce $w(n)$ or $\Delta w(n)$, which is feasible since either of the two needs be strictly positive if $W(\varepsilon, n) > W(\varepsilon, n-1) \geq \underline{w}(\varepsilon) \geq 0$. This, in turn, allows for a similar reduction of compensation at $n+1$ and ultimately at N, reducing equilibrium compensation costs.

IV. Asymmetric Contracts

Thus far, we have analyzed how symmetric contracts that condition on the firm's overall retention level n can reduce retention cost. In this section, we analyze the benefits of offering different (asymmetric) contracts to workers $i = 1, \ldots, N$, even though the workers are ex-ante identical. Our analysis of optimal asymmetric compensation design initially continues to assume that the firm seeks to retain all workers with probability one (full retention). We also discuss an extension to compensation design for asymmetric retention policies, in which workers are retained with different probabilities, as well as the optimality of such policies.

As we explain in Section II, the primary benefit of asymmetric full-retention contracts is that by offering compensation for which staying is an (iteratively) dominant strategy for some groups of workers, the remaining workers do not need to be compensated for the possibility that this group leaves, which makes them cheaper to retain. Existing work on asymmetric contracting, such as Winter (2004) and Halac, Kremer, and Winter (2020), studies the benefits of heterogeneous compensation *levels* for an exogenously given type of pay (e.g., fixed pay or call options). Instead, we focus on differences in compensation *structure*. That is, we study the optimal allocation of output-dependent and output-independent compensation across workers.

A. Optimal Asymmetric Compensation Design

To highlight how asymmetric contracts can help reduce coordination costs, we start by analyzing asymmetric contracts of the form $(w_i, \Delta w_i)_{i=1}^N$ that do not condition on the firm's overall retention level n. Given such asymmetric contracts, full retention is the unique equilibrium outcome if and only if staying is an iteratively dominant strategy for all workers and all shock realizations ε (Proposition 1). Without loss of generality, let the index i denote the *rank* of a worker. The highest-ranked worker ($i = 1$) stays no matter what other workers do, the lower-ranked worker $i = 2$ stays conditional on knowing that $i = 1$ has a dominant strategy to stay, and so on. Hence, the relevant participation constraints for worker i can be written as[22]

$$W_i(\varepsilon, i) = w_i + p(\varepsilon, i)\Delta w_i \geq \underline{w}(\varepsilon) \; \forall \varepsilon. \tag{11}$$

The firm's problem therefore is to design contracts $C_i = (w_i, \Delta w_i)$ that satisfy (11) for $i = 1, \ldots, N$ as well as the feasibility constraints $0 \leq \sum_i w_i \leq x$ and $0 \leq \sum_i \Delta w_i \leq \Delta x$ so as to minimize its expected equilibrium (full-retention) compensation costs or, equivalently, workers' expected rent.

[22] Given that at least $n = i - 1$ higher-ranked workers (those with lower index i) stay, the participation constraints $W_i(\varepsilon, n) = w_i + p(\varepsilon, n)\Delta w_i \geq \underline{w}(\varepsilon)$ for $n < i$ are irrelevant. Furthermore, the participation constraints for $n > i$ are implied by (11) since $W_i(\varepsilon, n) < W_i(\varepsilon, n + 1) < \cdots < W_i(\varepsilon, N)$, as $p(\varepsilon, n)$ is increasing in n.

Problem 2 (Optimal asymmetric full retention contracts)**:**

$$\min_{\{w_i,\Delta w_i\}_{i=1}^N \in \mathbf{C}^f} \sum_{i=1}^N \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \left[W_i(\varepsilon, N) - \underline{w}(\varepsilon)\right] dG(\varepsilon) \quad \textit{s.t. (11)}\ \forall i. \tag{12}$$

It follows from standard arguments that, at a solution to Problem 2, the participation constraint (11) of any worker i binds for at least one ε. Suppose, initially, that shocks are idiosyncratic ($\frac{\partial \underline{w}(\varepsilon)}{\partial \varepsilon} = 0$), such that (11) binds at $\underline{\varepsilon}$ for all i. This allows us to express worker i's output-dependent pay as a function of her output-independent pay according to $\Delta w_i = \frac{\underline{w} - w_i}{p(\underline{\varepsilon}, i)}$ and her expected rent as

$$\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \left[W_i(\varepsilon, N) - \underline{w}\right] dG(\varepsilon) = \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \left(\underline{w} - w_i\right) \left(\underbrace{\frac{p(\varepsilon, N)}{p(\underline{\varepsilon}, N)} \frac{p(\underline{\varepsilon}, N)}{p(\underline{\varepsilon}, i)}}_{= p(\varepsilon, N)/p(\underline{\varepsilon}, i)} - 1 \right) dG(\varepsilon). \tag{13}$$

Expression (13) illustrates that for any given $w_i < \underline{w}$, worker i earns a rent for two reasons. (i) First, since contracts cannot condition on n, the coordination problem leads the worker to "undervalue" output-dependent compensation relative to its equilibrium value when deciding whether to participate or not. Formally, this can be seen from the participation constraint (11) in which Δw_i is weighted with a probability of success of $p(\varepsilon, i)$, which is lower than the equilibrium full-retention probability, $p(\varepsilon, N)$, determining the firm's compensation costs, $W_i(\varepsilon, N) = w_i + p(\varepsilon, N)\Delta w_i$, in (12) for all $i < N$. Because of this misvaluation, the firm must offer a larger output-dependent component Δw_i to ensure participation than what the firm would need to offer absent the coordination problem. The effect on worker i's rent in (13) is captured by the ratio $p\left(\underline{\varepsilon}, N\right)/p\left(\underline{\varepsilon}, i\right) > 1$, which is decreasing in i. (ii) Second, just as with symmetric contracts, the worker also extracts rent because ε is not contractible. Specifically, given positive output-dependent pay, the difference between a worker's equilibrium on-the-job compensation $W_i(\varepsilon, N)$, and her outside option $\underline{w}$ increases in ε, which contributes to the worker's rent in (13) according to the ratio $p(\varepsilon, N)/p\left(\underline{\varepsilon}, N\right) > 1$, which is the same for all i.

As this discussion indicates, when risk is idiosyncratic, the firm can reduce the rent arising from both (i) and (ii) by paying workers with a higher share of output-independent pay. In fact, if resource constraints are slack, the firm can offer the *symmetric* fixed-pay only contract $(w_i, \Delta w_i) = \left(\underline{w}, 0\right)$ to all workers, under which neither the coordination problem nor the noncontractibility of ε lead to any cost, allowing the firm to extract all rent in line with our previous results.[23]

However, if this first-best solution is infeasible because resources in the low-cash-flow state, and thus the capacity for output-independent

[23] In fact, the cases in which $\frac{\partial}{\partial \varepsilon}\underline{w}(\varepsilon) \le 0$ combined with slack resource constraints ($x/N \le \underline{w}(\overline{\varepsilon})$) are the only ones in which the optimal contract (which does not condition on n) is symmetric.

compensation, are limited ($N\underline{w} > x$), minimizing workers' rent arising from the coordination problem makes it optimal for the firm to allocate output-independent compensation according to workers' rank. Specifically, the firm offers output-independent compensation up to $\underline{w}$ to worker $i = 1$ first, with workers $i = 2, 3, \ldots$ receiving fixed compensation only if resources are left. Formally, the optimal contract specifies output-independent compensation of $w_1 = \min\{x, \underline{w}\}$ and $w_i = \min\left\{x - \sum_{j=1}^{i-1} w_j, \underline{w}\right\}$ for $i \geq 2$ together with output-dependent compensation of $\Delta w_i = \frac{\underline{w} - w_i}{p(\underline{\varepsilon}, i)}$ for $i \geq 1$. The fundamental rationale underlying the optimality of this asymmetric compensation design is that the mispricing of any given level of output-dependent compensation is higher for higher-ranked workers (those with lower i) since their decision to stay must be independent of the decision of more workers. Optimal asymmetric compensation design addresses this by paying higher-ranked workers with a higher share of output-independent compensation (e.g., in the form of fixed pay, guaranteed bonuses, or retention bonuses).

This intuition cleanly extends to the case of systematic risk with $\frac{\partial}{\partial \varepsilon}\underline{w}(\varepsilon) > 0$ as long as workers' participation constraints in (11) bind either all "at the bottom" ($\underline{\varepsilon}$) or all "at the top" ($\overline{\varepsilon}$). In this case, minimizing the workers' rent arising from the noncontractibility of ε pushes toward the same compensation design (see Section III.C) for all workers independent of their rank.[24] Thus, the prediction that the firm will pay higher-ranked workers (lower i) with a higher share of output-independent pay again follows from the robust insight that they more strongly misprice output-dependent pay due to the coordination problem.

PROPOSITION 5: *Assume that contracts can condition on i (i.e., can be asymmetric) but cannot condition on the level of overall retention n. Then, the share of output-independent pay under the optimal contract is higher for higher-ranked workers if for all feasible contracts that achieve full retention, it holds either that* $\frac{\partial}{\partial \varepsilon}\left(W_i(\varepsilon) - \underline{w}(\varepsilon)\right) > 0$ *for all i or* $\frac{\partial}{\partial \varepsilon}\left(W_i(\varepsilon) - \underline{w}(\varepsilon)\right) < 0$ *for all i, that is, workers' participation constraints in (11) bind either all "at the bottom" (at* $\underline{\varepsilon}$*) or all "at the top" (at* $\overline{\varepsilon}$*). These conditions are always satisfied if shocks are idiosyncratic (*$\frac{\partial}{\partial \varepsilon}\underline{w}(\varepsilon) = 0$*) or systematic with* $\frac{\partial}{\partial \varepsilon}\underline{w}(\varepsilon) \leq 0$ *and the capacity for output-independent compensation is restricted,* $x < N\underline{w}$*, or if shocks are systematic with* $\frac{\partial}{\partial \varepsilon}\underline{w}(\varepsilon) \geq 0$ *and the capacity of output-dependent pay is restricted,* $\max_{n=1,\ldots,N}\left\{\frac{\partial p(\varepsilon, n)}{\partial \varepsilon}\Delta x\right\} < \frac{\partial}{\partial \varepsilon}\underline{w}(\varepsilon)$ *for all* ε.

To see the power of asymmetric compensation structure in lowering the firm's compensation costs, note that under the optimal asymmetric compensation design of Proposition 5, higher-ranked workers may earn lower rents than lower-ranked workers in equilibrium. We have already illustrated this

[24] If the regularity conditions we specify in Proposition 5 are not satisfied, workers' participation constraints might bind, for example, at $\underline{\varepsilon}$ for some ranks and at $\overline{\varepsilon}$ for others. In this case, the objective of minimizing the rent due to the noncontractibility of ε may push for different types of contracts for different ranks, which complicates an analytical characterization.

in the two-worker example in Section II.B. In this example with idiosyncratic risk ($\underline{w}(\varepsilon) = \underline{w}$), worker 1 is paid a fixed wage $(w_1, \Delta w_1) = (\underline{w}, 0)$ and has a dominant strategy to stay, while worker 2 is paid $(w_2, \Delta w_2) = \left(0, \frac{w}{p(\underline{\varepsilon},2)}\right)$ and has an iteratively dominant strategy to stay. Hence, the higher-ranked worker 1 extracts no rent, while worker 2 extracts a strictly positive rent for all $\varepsilon > \underline{\varepsilon}$. This insight stands in contrast to prior work on heterogeneous compensation, which takes compensation structure as exogenously given, since then higher-ranked workers, whose decision to stay must be independent of the decision of more workers, always receive higher equilibrium pay (e.g., Winter (2004)).

Having characterized the optimal design of asymmetric contracts that do not condition on n, we next consider their cost. In particular, we show that such contracts can resolve the coordination problem at no cost only if the optimal solution to the relaxed problem in which there are no coordination frictions heavily relies on output-independent pay. To see this, recall that all but the lowest-ranked worker $i = N$ misprice output-dependent pay due to the coordination problem (see participation constraint (11)). Thus, a necessary condition for the coordination problem to entail no cost is that all workers $i < N$ are paid a fixed wage. It follows that offering such a contract to workers $i < N$ can lead to the same aggregate compensation costs as in the relaxed problem without coordination frictions, that is, $\sum_{i=1}^{N} w_i = Nw^*(N)$ and $\sum_{i=1}^{N} \Delta w_i = N\Delta w^*(N)$, only if output-independent compensation, $w^*(N)$, in the relaxed problem is sufficiently large.[25]

PROPOSITION 6: *Assume that contracts can be asymmetric but cannot condition on n and take the solution, $(w^*(N), \Delta w^*(N))$, to the relaxed problem in which there are no coordination problems as given. Then, the firm can resolve the coordination problem at no cost if and only if (i) $w^*(N) \geq \frac{(N-1)}{N} \max_\varepsilon \underline{w}(\varepsilon)$, such that offering the fixed-wage contract $(w_i, \Delta w_i) = (\max_\varepsilon \underline{w}(\varepsilon), 0)$ to workers $i < N$ is feasible and (ii) $W^*(\varepsilon, N) \geq \frac{1}{N}\left((N-1)\max_\varepsilon \underline{w}(\varepsilon) + \underline{w}(\varepsilon)\right)$ for all ε such that offering worker N the contract $(w_N, \Delta w_N) = \left(Nw^*(N) - (N-1)\max_\varepsilon \underline{w}(\varepsilon), N\Delta w^*(N)\right)$ satisfies this worker's participation constraint.*

B. Asymmetric Versus Symmetric Contracts

In Sections III and IV.A, we show that optimal contracts that can condition on *either* the firm's overall retention level n *or* workers' identity i can mitigate the coordination problem among workers, and we characterize the conditions under which they can fully resolve this problem. Below, we first compare the compensation costs associated with these contracts, and we discuss when

[25] Note that it is without loss of generality to restrict attention to symmetric contracts in the relaxed problem. To see this, consider any asymmetric compensation scheme conditional on full retention $(w_i(N), \Delta w_i(N))$ for $i = 1, \ldots, N$. Then, a symmetric contract paying each worker the average of all asymmetric contracts $\left(\frac{1}{N}\sum_{i=1}^{N} w_i(N), \frac{1}{N}\sum_{i=1}^{N} \Delta w_i(N)\right)$ can retain all workers at the same aggregate compensation cost.

combining contract asymmetry (i.e., contracting on workers' identity i) with contracting on the firm's retention level n can further lower coordination costs.

Contracting on the level of retention n tends to be more cost-efficient than contracting on workers' identity i because it allows the firm to better disentangle workers' coordination problem from the noncontractibility of ε. In particular, the firm can solve the coordination problem *off-equilibrium* by offering payments $\big(w(n), \Delta w(n)\big)$ for out-of-equilibrium retention levels $n < N$ that differ from the full-retention payments $\big(w(N), \Delta w(N)\big)$ paid to workers in equilibrium. Doing so fully eliminates the cost arising from the coordination problem if the monotonicity-in-retention constraint remains slack (Proposition 3 and Lemma 2).

By contrast, asymmetric contracts must resolve the coordination problem *on-equilibrium*. In particular, ensuring that staying is an (iteratively) dominant strategy requires promising all but one worker equilibrium compensation under which staying is optimal even if a subset of workers leaves (Proposition 5). Thus, if workers $i < N$ are paid with output-dependent compensation, the firm will always pay these workers higher compensation than it would in the absence of coordination frictions (Proposition 6). Yet, because these contracts do not condition on n, they trivially satisfy the monotonicity-in-retention constraint. Taken together, these results imply that symmetric contracts that condition on n are cheaper for the firm, unless binding feasibility constraints make the monotonicity-in-retention constraint too costly to satisfy. To make this generic result more tangible, the proof of the following lemma provides a precise condition for the case of $x = 0$, which can be solved in closed form.

LEMMA 3: *Optimal symmetric contracts that can condition on the firm's retention level n imply lower compensation costs than optimal asymmetric contracts that cannot condition on n unless the shadow cost of the monotonicity-in-retention constraint under optimal symmetric contracting is sufficiently high.*

Combining asymmetric contracting (contracting on i) with contracting on the firm's retention level n always lowers the firm's retention costs if contracting on n alone cannot resolve the coordination problem at no cost. It does so by combining the benefits of dilution with those of requiring that staying only be an iteratively dominant strategy instead of a dominant strategy. Requiring only iterative dominance reduces the set of relevant (out-of-equilibrium) retention scenarios for lower-ranked workers since they can be sure that higher-ranked workers are always staying. This means that there are fewer relevant participation and monotonicity-in-retention constraints for lower-ranked workers, which leads to lower compensation costs if these constraints bind under the optimal symmetric contract.

PROPOSITION 7: *Suppose that the optimal symmetric contract that conditions on n entails positive coordination costs. Then, there always exists an asymmetric contract that conditions on n that strictly reduces the firm's expected compensation costs relative to the optimal symmetric contract.*

To illustrate the intuition and show what this optimal contract looks like, suppose that $N = 2$ and that coordination is costly under the optimal symmetric full-retention contract from Section II, that is, $E[W(\varepsilon, N)] > E[W^*(\varepsilon, N)]$. Then, from Proposition 7, the firm can strictly reduce expected compensation costs relative to optimal symmetric contracting by offering different contracts to the two workers. Specifically, assume that the firm offers the optimal symmetric contract to worker 1. Since staying is a dominant strategy for worker 1 under this contract, the firm only needs to ensure that staying is an iteratively dominant strategy for worker 2 by offering her a feasible contract that pays her more in expectation than the outside option conditional on $n = N = 2$. Since the optimal contract $(w^*(N), \Delta w^*(N))$ from the relaxed problem satisfies this condition and is feasible, full retention can be achieved by offering this contract to worker 2 for all retention levels. This reduces the average per-worker compensation cost to $\frac{1}{2}(\mathrm{E}[W(\varepsilon, N)] + \mathrm{E}[W^*(\varepsilon, N)]) < E[W(\varepsilon, N)]$. Under a mild condition that we discuss in the Internet Appendix, this asymmetric contract represents the optimal full-retention contract.

C. Asymmetric Retention

So far, we have considered optimal contracts under which the firm retains all workers with probability one. While we assume that such full retention is efficient, it may force the firm to leave substantial rents to workers, which might lead the firm to prefer contracts that do not achieve full retention for all shock realizations for all workers. Analyzing whether the firm can do better with contracts that do not ensure full retention faces the problem that such contracts result in a multiplicity of equilibria if they satisfy condition (3). However, our model can be easily extended to tackle this problem by using a standard global games refinement under which workers observe informative but imperfect signals of the shock realization ε (see, e.g., Morris and Shin (2003)). We relegate a complete formal analysis to the Internet Appendix (Propositions I.A1 to I.A7). Here, we provide an informal discussion of the main economic insights, illustrated based on the case of idiosyncratic shocks.

When shocks are idiosyncratic, workers' expected on-the-job pay increases in the now unobserved productivity shock ε, while their outside option remains constant, $\frac{\partial}{\partial \varepsilon}\left(W(\varepsilon, n) - \underline{w}\right) \geq 0$. Accordingly, staying becomes more attractive for a given worker if she observes a higher signal since a higher signal is indicative of a higher productivity shock. Thus, by standard arguments, equilibria are characterized by a cutoff rule: any given worker wants to stay following high signals above a cutoff ε^* and leave following low signals below that cutoff. Since signals are drawn independently across workers, such interior cutoff rules imply that any number of workers $0 \leq n \leq N$ may stay in equilibrium, even if the cutoffs, and therefore the ex ante retention probabilities, are the same for all workers. For this reason, we refer to a set of such cutoffs (one for each worker) as a *partial retention* policy. We refer to the case in which cutoffs differ across workers as an *asymmetric retention* policy. In the following, we describe the primary features of optimal compensation design for a given symmetric- or asymmetric-retention policy and subsequently discuss the optimality of asymmetric retention.

Optimal Asymmetric Retention Contracts. Optimal partial retention contracts bear many similarities to optimal full-retention contracts. In particular, in determining the optimal mix of output-dependent and output-independent pay, the firm tries to match each worker's on-the-job pay to their outside options as closely as possible, just as it does for optimal full-retention contracts (Proposition 4).[26] Furthermore, optimal contracts will be dilutable. Indeed, since any retention level now can arise in equilibrium—with n increasing in workers' signals—contracting on n allows the firm to indirectly contract on the workers' private signals. Thus, the firm can extract more rent by making partial retention contracts dilutable (Proposition IA1).

The main novel predictions for optimal compensation design arise when the retention policy is asymmetric. We illustrate these predictions with a case in which the firm seeks full retention of one group of workers ("group 1" with cutoff $\varepsilon_1^* = \underline{\varepsilon}$) but is willing to accept that the remaining workers leave with positive probability ("group 2" with cutoff $\varepsilon_2^* > \varepsilon_1^*$). With idiosyncratic risk, such an asymmetric retention policy can be optimal only if resources in the low-cash-flow states are limited and the firm cannot resolve the coordination problem at no cost by offering all workers a fixed-wage contract perfectly matching their outside option. In this case, using the firm's limited resources to offer group 1 more output-independent compensation to lower their rent mechanically implies that the firm cannot offer as much output-independent compensation to group 2, resulting in higher rent for group 2.

The key observation is that these two effects do not offset each other perfectly. The reason is that output-dependent pay is worth less for signals at the lower cutoff ε_1^* (relevant for retaining workers for group 1) than at the higher cutoff $\varepsilon_2^* > \varepsilon_1^*$ (relevant for group 2) since the probability of receiving such pay is lower, $p\left(\varepsilon_1^*, n\right) < p\left(\varepsilon_2^*, n\right)$. Instead, the value of output-independent pay is the same for all signals. Because of this difference, the firm should allocate its limited resources to promise higher output-independent compensation to group 1 so as to lower the (relatively more expensive) output-dependent compensation it needs to offer this group. Hence, the firm pays workers that it retains with a higher probability with a higher share of output-independent pay (Proposition I.A2). As a result, these workers do not necessarily extract more rent. A simple example illustrating this is when workers from group 1 are paid a fixed-wage contract of $\underline{w}$.

These features of optimal asymmetric retention contracts are reminiscent of our analysis of asymmetric full-retention contracts that do not condition on n, where higher-ranked workers are also paid with a higher share of output-independent pay and do not necessarily extract more rent (Section IV.A). Yet,

[26] Specifically, if shocks are idiosyncratic, the firm will aim to induce the retention cutoff ε^* with the lowest proportion of output-dependent pay subject to feasibility constraints. The only novel element relative to the full-retention benchmark is that the firm will also do so for retention levels $n < N$ that can now also arise in equilibrium.

the economic forces behind these predictions are not the same. When cutoffs are the same—as they are in the case of a full-retention policy—but contracts cannot condition on n, the driving force is that the coordination problem leads to higher mispricing of output-dependent pay for higher-ranked workers, $p(\underline{\varepsilon}, i) < p(\underline{\varepsilon}, N)$. By contrast, the driving force in the presently considered setting, in which contracts can condition on n but the retention policy is asymmetric, is that the shock ε is noncontractible. In particular, the cost of output-dependent compensation to the seller is determined by how it is valued by the buyer at the respective cutoff, $p(\varepsilon_1^*, n) < p(\varepsilon_2^*, n)$ for $\varepsilon_1^* < \varepsilon_2^*$.

This difference in the key friction driving optimal asymmetric compensation design between the two settings becomes apparent when risk is systematic, with workers' outside options being much more sensitive to shocks than firm output. While the qualitative compensation design predictions do not change relative to the case in which shocks are idiosyncratic if the main friction is the coordination problem (Proposition 5), the predictions may reverse if the main friction is the noncontractibility of ε. In particular, while in this case, the firm will also allocate its limited resources to the group of workers it wants to retain with higher probability, the noncontractibility of ε calls for offering this group compensation with a higher proportion of output-dependent pay (Proposition I.A5).[27]

Optimality of Asymmetric Retention. We close this section by briefly commenting on the optimality of asymmetric retention. The main insights are simple to illustrate in an example of two workers and idiosyncratic risk. In this case, if offering both workers a symmetric fixed-wage contract that matches their outside option is feasible (i.e., if $x \geq 2\underline{w}$), it is optimal because it maximizes efficiency and allows the seller to extract all rent from the workers. However, if the firm is resource-constrained (i.e., $x < 2\underline{w}$), it faces a trade-off between rent extraction and efficiency.

This trade-off can be mitigated by a *combination* of asymmetric retention and optimal compensation design. To see this, note that for any symmetric retention policy, both workers must be paid at least in part with output-dependent compensation, and thus earn a rent, even if dilution provisions resolve the coordination problem at no cost. If this rent is large under full retention, the firm may find it optimal to sacrifice some efficiency under the optimal symmetric contract by choosing an interior cutoff ε^* to extract more rent. In this case, if $x \geq \underline{w}$, the firm can improve on this symmetric contract by offering worker 1 a fixed-wage contract $\underline{w}$ and worker 2 a contract for which her retention cutoff is unchanged at ε^*. This asymmetric retention contract makes the firm better off since it can extract the entire efficiency increase from retaining worker 1 with probability one while also ensuring that worker 2 faces no coordination problem, making that worker at least weakly cheaper to retain. We formalize this argument in Proposition I.A3 in

[27] Since outside options are binding "at the top," the group retained with a higher probability (group 1) will now be the one with a higher cutoff. At this cutoff, output-dependent pay is worth more than at the cutoff for group 2 since $p(\varepsilon_1^*, n) > p(\varepsilon_2^*, n)$ for $\varepsilon_1^* > \varepsilon_2^*$.

the Internet Appendix. Complementing our analysis, Luo and Yang (2024) provide an in-depth treatment of how to optimally differentiate across agents in coordination games with optimal security design (but no contracting on n) in a model in the spirit of Sakovics and Steiner (2012).

V. Implementation and Empirical Implications

Our focus on inefficient contagious turnover that firms try to prevent is likely to be relevant for firms that rely heavily on well-functioning teams of high-skilled workers with complementary skills. High-growth startups, for which "an A team with a B idea is more important than a B team with an A idea," are a prime example since their human capital is usually less standardized than in mature public firms (Rajan (2012)). Other examples include firms in consulting, advisory, investment banking, legal services, and private equity partnerships, in which departures of key workers can be highly disruptive to the team and reduce everyone's productivity.[28] Large firms organized around teams, as in our example of Credit Suisse in the introduction, are also affected. Studies of turnover contagion in large firms show that it is two-to-three times more likely when they are organized around smaller teams (Derler et al. (2023)). Furthermore, contagious turnover is more likely if firm productivity remains low long after the departure of a single worker, which might be the case if certain skills are acquired on the job only slowly or are hard to find in the market. Proxies for how difficult it is to replace workers could be constructed based on data about how quickly firms can fill job vacancies.

Degree of Dilution. Our key compensation design prediction is that whenever firms vulnerable to worker runs offer output-dependent compensation, they should make workers' compensation dilutable. The reason is that dilutable compensation helps to cost-effectively improve collective retention. Firms benefit from offering dilutable pay, as it helps to both counter workers' incentives to leave when others are leaving and lower workers' rents when retention levels are high. For the degree of dilution, the following implication holds:

IMPLICATION 1: *When firms relying on teams of hard-to-replace skilled workers pay these workers with output-dependent pay, they should optimally make workers' compensation dilutable. The degree of dilution is higher when:*

(i) firm performance is more sensitive to retention and
(ii) the share of workers' output-dependent compensation is higher—such as when firms are more cash-constrained or workers' outside options are positively correlated with (or become more sensitive to) systematic shocks positively affecting firm output.[29]

[28] In general, human-capital-intensive industries, such as finance (e.g., Credit Suisse in 2021), tech (e.g., Infosys in 2014), and legal services (e.g., Sedgwick in 2017), are often affected by inefficient collective turnover. See, for example, "Leaving the dream: Infosys battles worker exodus," Reuters, May 11, 2014.

[29] In our model, the optimal mix of output-dependent and output-independent compensation is derived from the firm's resource constraints and the friction that the shocks affecting firm

As we argue above, dilution is a feature of standard (time-vesting) equity compensation. Thus, Implication 1 can help inform discussion of why dilutable equity-based pay is commonly used for retention purposes (Aldatmaz, Ouimet, and Van Wesep (2018), Jochem, Ladika, and Sautner (2018), Hochberg and Lindsey (2010)) despite arguments that deferred fixed pay better insulates workers against risk beyond their control (Murphy (2003), Lazear (2004)). The use of equity-based pay for retention purposes is especially notable in startups, which rely extensively on such compensation (Hand (2008)). Such evidence squares well with our model, as startups not only rely crucially on keeping teams of workers together but also on diluting employees' percentage equity ownership in states of the world in which these startups are doing well and employees are unlikely to leave.[30] We should emphasize again that dilution is not specific to equity-based pay since most output-dependent compensation can be made dilutable. As we discuss below, other examples include bonus pools at the division or team level or retention bonuses. Such compensation can also disentangle dilution from firm size and tie dilution instead to the size of teams around which firms are organized.

Dilution Patterns. The optimal way to make compensation dilutable depends on whether technological shocks and retention are complements or substitutes in firms' production (Corollary 2). In the case of complements, positive technological shocks increase the probability of firm success more for high retention levels. This case could arise if shocks positively affect labor productivity in the sense of increasing the productivity contribution per worker or if a technological innovation requires a certain level of workforce capacity to be successfully implemented. In this case, dilution is achieved primarily by making output-dependent pay decreasing in retention. Instead, the case of substitutes captures the opposite scenario, in which the firm benefits most from technological shocks at low levels of retention. This could be the case following an innovation in a technology that can substitute for labor, thus lowering the benefit of retention. In this case, dilution is achieved primarily by making output-independent pay decreasing in retention.

IMPLICATION 2: *Dilution provisions are more likely to be tied to (i) output-dependent pay (i.e., $\Delta w(n)$ is decreasing in n) if technological shocks increase firm productivity more for higher retention levels and (ii) output-dependent pay (i.e., $w(n)$ is decreasing in n) if these shocks increase productivity more if retention is low.*

Implementation. Consider more closely the case of complements. A dilution pattern in which both $w(n)$ and $\Delta w(n)$ are decreasing in n (Panel A.3 of Figure 2) can be implemented using standard time-vesting equity, as seen

productivity and workers' outside options are not contractible. One could consider additional reasons for offering more output-dependent pay, such as incentive reasons, or less output-dependent pay, such as providing insurance to risk-averse workers.

[30] Specifically, the firm may issue shares to new employees or to nonemployee owners or investors in new funding rounds. Alternatively, the firm may issue warrants that allow founders or investors to purchase additional equity in high-cash-flow states.

in the simple example following Definition 1.[31] If, instead, only the output-dependent component, $\Delta w(n)$, is decreasing in n (Panel A.2), the firm can offer stock options instead of equity. Alternatively, the firm may offer equity with performance-vesting along with time-vesting provisions, which is also highly common in practice (Edmans, Gabaix, and Jenter (2017), Bettis et al. (2018)). As an alternative to equity-based pay, firms can make performance bonuses dilutable. For example, firms can set aside a bonus pool as a percentage of a division's net revenues to be shared among employees from that division at the end of a prespecified period. The more employees stay until the end, the higher the expected revenue generated by the division, but the lower the share of the pool each worker receives.

Corollary 2 further predicts that if complementarity is high, firms should optimally make output-dependent pay, $\Delta w(n)$, decreasing in n and output-independent pay, $w(n)$, increasing in n (Panel A.1 of Figure 2). This contract can be implemented through equity buyback agreements. Such contracting arrangements are common, for example, in law firms. When partners in a law firm leave, they usually sell their equity back to the firm. Firms pay for that equity either through their working capital or by taking out a credit line, with the partners typically individually guaranteeing the repayment of the facility. Hence, a departure by a partner results in a higher $\Delta w(n)$ and a lower $w(n)$ for remaining partners. Similarly, when key employees leave a startup, the firm buys back their unvested equity (typically at par) and sometimes also their vested equity (at its fair market value at the grant date or a price specified in the employee's employment agreement). [32]

If there is no complementarity between shocks and retention (Panel B of Figure 2), then making only the output-independent component dilutable is possible via retention bonuses offered to remaining employees when others leave. Finally, if retention and shocks act as substitutes, it is optimal to make $w(n)$ decreasing in n and $\Delta w(n)$ increasing in n. Such a pattern, which effectively shifts output-dependent compensation, $\Delta w(n)$, to output-independent compensation, $w(n)$ (Panel C of Figure 2), could be achieved by stipulating that retention bonuses replace performance bonuses for low retention levels.[33]

Asymmetric Compensation. Another lever that firms can use to reduce the cost of mitigating worker runs is to compensate (identical) workers differently. Since, in practice, workers are rarely identical, this means that even marginal differences among workers can lead to large differences in compensation

[31] Equity can be presented by a fraction $\nu(n)$ of firm cash flows, implying that $w(n) = \nu(n)x$ and $\Delta w(n) = \nu(n)\Delta x$. As we illustrate in our example on page 14, time-vesting equity makes $\nu(n)$ decreasing in n, implying that both $w(n)$ and $\Delta w(n)$ decrease in n.

[32] See "Startup Employee Alert: Can Your Company Take Back Your Vested Shares?," Forbes, January 10, 2018.

[33] For example, Whiting Petroleum's retention bonus agreement states that "*As a condition to receiving the Retention Bonus you hereby (i) waive any and all participation in any annual bonus plan established by the Company for the 2020 calendar year and (ii) agree to forfeit and terminate any equity-based award previously made by the Company to you during calendar year 2020.*"

design. There is abundant anecdotal evidence of firms offering asymmetric pay in practice. For example, as Winter (2004) discusses, it is standard in investment banking, law, and consulting firms to assign titles such as Associate and Senior Associate or Partner and Principal, which map into differences in pay but not in authority. While Winter's model suggests that workers are assigned higher ranks associated with higher pay levels when the firm seeks to motivate these workers regardless of whether their coworkers are motivated, in our paper, the different compensation design associated with higher ranks ensures that these workers can be retained with higher probability or independent of the retention of others. Another example of asymmetric compensation is firms trying to ensure the retention of a critical mass of workers by targeting key employees with retention bonuses. For instance, in June 2021, the financial press extensively reported that Credit Suisse was targeting only a subset of its employees with retention bonuses to prevent the mass exodus of staff concerned with being "the last man standing." Such targeting of select workers with safer compensation, like retention bonuses, is consistent with our predictions in Proposition 5. As we further show, the benefit of offering asymmetric compensation is particularly relevant for resource-constrained firms.

IMPLICATION3 (i): *Firms can lower the cost of preventing worker runs by offering (identical) workers different types of compensation. In particular, firms can benefit from making the compensation of some employees safer by offering them a higher share of output-independent pay (fixed pay or retention bonuses) while paying the remaining employees with a higher share of output-dependent pay (equity or bonus payments). Resource-constrained firms are especially likely to offer such asymmetric compensation. (ii) If firms seek to retain some workers with higher probability, firms highly exposed to idiosyncratic risk optimally pay these workers with a higher share of output-independent pay. By contrast, if workers' outside options are more sensitive than firm output to common shocks, targeted workers will receive a higher share of dilutable performance- or equity-based pay.*

Tests of Implications 1 and 3 will have to rely on determining the sensitivity of workers' outside options to systematic shocks. This requires identifying the most attractive outside employment opportunities for the firm's skilled workforce. Such opportunities can typically be found in closely related firms operating in the same industry or requiring a similar skill set. [34]

[34] Identifying such similar firms is crucial to adequately capturing the notion of "relative shock sensitivity (ρ)" (described in the Internet Appendix), which corresponds to a ratio of sensitivities to a common factor. This ratio is related but does not immediately map into proxies of comovement between a firm and its local peers in Kedia and Rajgopal (2009). To identify closely related firms to construct such a proxy, one may use Hoberg and Phillips (2016) network similarity scores. A proxy for the equity-based pay offered by firms could be the proportion of equity-based pay as measured in Bergman and Jenter (2007).

VI. Conclusion

The high quit rate at human capital-intensive firms and the trillion of dollars in costs related to rehiring and training new workers make improving retention a first-order priority for firms. While avoiding the departure of individual workers is certainly costly, a further substantial threat is that turnover will become contagious: fearing that the departure of skilled workers will erode firm performance, workers begin leaving in droves. This risk of "worker runs" is particularly relevant when workers' compensation depends on firm performance and thus on its ability to retain productive employees, such as when workers are paid with performance bonuses or equity. In our model, firms may use such compensation to better match workers' on-the-job pay to their outside options or because cash constraints prevent them from offering high-enough fixed pay.

When workers are paid with output-dependent pay, firms can mitigate the risk of runs by making compensation dilutable. Typical examples of dilutable pay include (time- and performance-) vesting equity and stock options, bonus-sharing schemes, or retention bonuses. The principle of how dilutable compensation counters contagious turnover is simple to illustrate with time-vesting equity: because workers forgo their deferred equity compensation when leaving, the remaining workers' percentage equity stake increases. Intuitively, while the size of the pie may decrease with the loss of human capital, remaining workers' share of the pie increases, reducing their incentives to leave. Conversely, diluting workers' percentage equity share is optimal when the firm is doing well and overall retention is high since, then, the value of equity is high. Taken together, making compensation dilutable lowers the strategic complementarity in workers' decisions to stay or leave and can allow the firm to reduce workers' rents.

An alternative solution to the worker run problem is offering different compensation contracts to even ex-ante identical workers. The idea is to ensure that a critical mass of workers is always retained, as this reduces the risk of contagious turnover. In practice, the optimality of compensating workers differently means that small differences across workers could lead to large differences in compensation design. Notably, compensating workers differently does not necessarily require paying a subset of workers strictly more. Instead, we show how this can be achieved by changing the structure of pay. While introducing asymmetries in contracting is, in principle, less cost-efficient than offering dilutable compensation, combining the two tools can lower retention costs when feasibility constraints prevent the firm from optimally designing dilutable compensation. Overall, our paper shows that mitigating collective turnover poses additional challenges relative to individual turnover and therefore requires different compensation design solutions. Notably, the tools we propose—making compensation dilutable or offering asymmetric compensation—can easily be adapted to other objectives of compensation design and can be easily implemented with well-understood compensation provisions.

Initial submission: June 29, 2022; Accepted: January 30, 2024
Editors: Stefan Nagel, Philip Bond, Amit Seru, and Wei Xiong

Appendix: Proofs

PROOF OF PROPOSITION 1: Fix C_i for all $i \in I$ and denote the set of workers who have an iteratively dominant strategy to stay by $S \subseteq I$ and the set of workers with an iteratively dominant strategy to leave by $L \subseteq I$, where $N^s = \left|S\right|$ and $N^l = |L|$.[35] By definition, in any equilibrium of the workers' coordination game at $t = 1$, the N^l workers in L leave and the N^s workers in S stay. Thus, the equilibrium of this game is unique if $N^l + N^s = N$. We show that whenever $N^l + N^s < N$, there is a multiplicity of equilibria involving a worker-run equilibrium in which a subset of workers $R \subseteq I \backslash \left\{L \cup S\right\}$ leaves despite the fact that all workers $i \in I \backslash \left\{L \cup S\right\}$ would be better off if they stayed. To see this, note that (3) needs to hold for all workers $i \in I \backslash \left\{L \cup S\right\}$ who do not have an iteratively dominant strategy. But this directly implies that there are at least two equilibria: one in which all $N - N^l - N^s$ workers in $I \backslash \left\{L \cup S\right\}$ stay with the corresponding correct beliefs, in which case each worker realizes a payoff of $W_i(\varepsilon, N - N^l)$, and one in which they all leave and forgo their compensation in favor of their outside option $\underline{w}(\varepsilon)$. From condition (3), each worker $i \in I \backslash \left\{L \cup S\right\}$ clearly prefers the former equilibrium. In fact, it follows directly from Assumption 1 that everyone receives their maximally achievable equilibrium payoff in the equilibrium in which all $i \in I \backslash \left\{L \cup S\right\}$ stay. □

PROOF OF LEMMA 1: Consider a point $\varepsilon_n \in \left[\underline{\varepsilon}, \overline{\varepsilon}\right]$ at which $W_i\left(\varepsilon_n, n-1\right) \geq W_i\left(\varepsilon_n, n\right)$. We then have

$$\begin{aligned} w_i(n-1) + \widehat{p}\Delta w_i(n-1) &\geq w_i(n-1) + p(\varepsilon_n, n-1)\Delta w_i(n-1) \\ &\geq w_i(n) + p(\varepsilon_n, n)\Delta w_i(n) \geq w_i(n) + \widehat{p}\Delta w_i(n), \end{aligned}$$

where the first inequality is strict if $\widehat{p} \in \left(p\left(\varepsilon_n, n-1\right), p\left(\varepsilon_n, n\right)\right]$, and the last inequality is strict if $\widehat{p} \in \left[p\left(\varepsilon_n, n-1\right), p\left(\varepsilon_n, n\right)\right)$, proving that the contract is dilutable at ε_n. □

PROOF OF PROPOSITION 2: If $w\left(N\right)$ is a feasible fixed-wage contract, then setting $w\left(n\right) = w\left(N\right)$ is also feasible, since the resource constraint $w\left(n\right) \leq \frac{x}{n}$ is tightened in n. To see the first statement of Proposition 2, note that the optimal contract must satisfy the participation constraint $W\left(\varepsilon, n\right) \geq \underline{w}\left(\varepsilon\right)$ for all retention levels n. Furthermore, since $\left(w^*(N), \Delta w^*(N)\right)$ is an optimal solution to the relaxed problem, it will satisfy workers' full-retention participation

[35] Workers for which $W_i(\varepsilon, 1) \geq \underline{w}(\varepsilon)$ have a strictly dominant strategy to stay given our tie-breaking assumption, and we denote their number by N_1^s. Then, if $N_1^s > 0$, also workers for which $W_i(\varepsilon, N_1^s + 1) \geq \underline{w}(\varepsilon) > W_i(\varepsilon, 1)$ stay in any equilibrium since they can be sure that N_1^s workers stay. We denote the number of workers with such an "iteratively dominant strategy" after one step by N_2^s. Proceeding in this fashion until there is no worker for which $W_i(\varepsilon, N_1^s + \cdots + N_M^s + 1) \geq \underline{w}(\varepsilon) > W_i(\varepsilon, N_1^s + \cdots + N_{M-1}^s + 1)$, that is, $N_{M+1}^s = 0$, we obtain the total number of workers who have an iteratively dominant strategy to stay $N^s = N_1^s + \cdots + N_M^s$. The set of workers who have an iteratively dominant strategy to leave can be characterized similarly.

constraint, $W^*(\varepsilon, N) \geq \underline{w}(\varepsilon)$, with equality for at least some shock realization ε_N. Hence, satisfying the remaining participation constraints for $n < N$ at ε_N requires that $W(\varepsilon_N, n) \geq \underline{w}(\varepsilon_N) = W(\varepsilon_N, N)$. From Lemma 1, this relation implies that the contract is dilutable at (ε_N, N) and, by continuity of $p(\varepsilon, n)$ in ε, in the immediate neighborhood of ε_N whenever $\Delta w^*(N) > 0$. The second statement of Proposition 2 follows immediately from the same arguments and our discussion in Section II.B. □

PROOF OF PROPOSITION 3: Abstracting from feasibility constraints, consider the relaxed problem in which we abstract from all interim participation constraints for retention levels $n < N$:

$$\big(w^*(N), \Delta w^*(N)\big) := \arg \min_{w(N), \Delta w(N)} \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \big[w(N) + p(\varepsilon, N)\Delta w(N) - \underline{w}(\varepsilon)\big] dG(\varepsilon) \quad \text{(A1)}$$
$$s.t.\; w(N) + p(\varepsilon, N)\Delta w(N) - \underline{w}(\varepsilon) \geq 0.$$

We now show that there exists a contract C^* with $\big(w(N), \Delta w(N)\big) = \big(w^*(N), \Delta w^*(N)\big)$ that implements full retention as a dominant strategy and hence solves the firm's unrestricted problem at the same compensation costs. In particular, this contract sets the out-of-equilibrium promises $\big(w(n), \Delta w(n)\big)_{n=1}^{N-1}$ such that $W(\varepsilon, n) = W(\varepsilon, N) = w^*(N) + p(\varepsilon, N)\Delta w^*(N)$ for all (ε, n). To construct such a contract, note that

$$\frac{\partial W(\varepsilon, n)}{\partial \varepsilon} = \frac{\partial p(\varepsilon, n)}{\partial \varepsilon} \Delta w(n) = \beta'(\varepsilon)\gamma(n)\Delta w(n), \quad \text{(A2)}$$

where we have used (1). Hence, requiring that $\partial W(\varepsilon, n)/\partial \varepsilon = \partial W(\varepsilon, n-1)/\partial \varepsilon$ for all ε boils down to

$$\Delta w^*(n) = \frac{\partial p\big(\varepsilon, n'\big)/\partial \varepsilon}{\partial p(\varepsilon, n)/\partial \varepsilon} \Delta w^*(n') = \frac{\gamma(n')}{\gamma(n)} \Delta w^*(n') \text{ for any } n \neq n'.$$

From this, it follows immediately that the contract is given by (9) and (10), respectively, where we use the fact that the contract under full retention is given by C^*. Since the suggested contract is monotonic in retention by construction, that is, satisfies Assumption 1, it is feasible if and only if $0 \leq w^*(n) \leq x/n$ and $0 \leq \Delta w^*(n) \leq \Delta x/n$ for all n. Furthermore, by Lemma 1, this contract is dilutable if $\Delta w^*(N) > 0$.

It remains to show that the suggested contract is uniquely optimal for any given interior solution to the relaxed problem. To see this, note that dominant-strategy implementation (6) together with the monotonicity-in-retention constraint (Assumption 1) requires condition (7) to hold. Note next that optimality of $\big(w^*(N), \Delta w^*(N)\big)$ from (A1) implies that $W(\varepsilon, N) = \underline{w}(\varepsilon)$ for at least one $\varepsilon = \varepsilon_0$. Hence, at ε_0, (7) has to hold with equality. If $\varepsilon_0 \in \big(\underline{\varepsilon}, \overline{\varepsilon}\big)$, (7) further requires that $\partial W(\varepsilon_0, n)/\partial \varepsilon = \partial W(\varepsilon_0, n-1)/\partial \varepsilon$ for all n, since $W(\varepsilon_0, n)$ must then be tangent to $\underline{w}(\varepsilon_0)$ for all n. Hence, from (A2), we must have $\partial W(\varepsilon, n)/\partial \varepsilon = \partial W(\varepsilon, n-1)/\partial \varepsilon$ for all (ε, n). But we then obtain that $W(\varepsilon, n) = W(\varepsilon, n-1)$

has to hold for all (ε, n). Hence, if $W(\varepsilon, N) \geq \underline{w}(\varepsilon)$, then also $W(\varepsilon, n) \geq \underline{w}(\varepsilon)$ for all $n < N$. If instead $\varepsilon_0 \notin (\underline{\varepsilon}, \overline{\varepsilon})$, optimality implies that $W(\varepsilon, N) = \underline{w}(\varepsilon)$ holds at both $\varepsilon = \underline{\varepsilon}$ and $\varepsilon = \overline{\varepsilon}$. But if (7) holds with equality at both corners, it again follows from (A2) that $W(\varepsilon, n) = W(\varepsilon, n-1)$ has to hold for all (ε, n). □

PROOF OF COROLLARY 1: Since $W(\varepsilon, n)$ is constant in retention, it is convenient to express from $W(\varepsilon, n) = W^*(\varepsilon, N)$

$$\Delta w^*(n) = \frac{\gamma(N)}{\gamma(n)} \Delta w^*(N)$$

$$w^*(n) = w^*(N) + \left(\alpha(N) - \alpha(n)\frac{\gamma(N)}{\gamma(n)}\right)\Delta w^*(N).$$

Substituting into $\frac{\left|\widehat{W}(\varepsilon,n)-\widehat{W}(\varepsilon,n-1)\right|}{\widehat{W}(\varepsilon,n)}$, we obtain

$$\frac{\left|w^*(n) - w^*(n-1) + \widehat{p}[\Delta w^*(n) - \Delta w^*(n-1)]\right|}{w^*(n) + \widehat{p}\Delta w^*(n)}$$

$$= -\frac{\left((\widehat{p} - \alpha(n))\frac{\gamma(N)}{\gamma(n)} + (\alpha(n-1) - \widehat{p})\frac{\gamma(N)}{\gamma(n-1)}\right)\Delta w^*(N)}{w^*(N) + \left(\alpha(N) - \alpha(n)\frac{\gamma(N)}{\gamma(n)}\right)\Delta w^*(N) + \widehat{p}\frac{\gamma(N)}{\gamma(n)}\Delta w^*(N)} \tag{A3}$$

$$= \frac{(\alpha(n) - \widehat{p})\frac{\gamma(N)}{\gamma(n)} + (\widehat{p} - \alpha(n-1))\frac{\gamma(N)}{\gamma(n-1)}}{\frac{w^*(N)}{\Delta w^*(N)} + \left(\alpha(N) - \alpha(n)\frac{\gamma(N)}{\gamma(n)}\right) + \widehat{p}\frac{\gamma(N)}{\gamma(n)}}, \tag{A4}$$

where the negative sign in (A3) follows from the fact that $\widehat{W}(\varepsilon, n) \leq \widehat{W}(\varepsilon, n-1)$ if the contract is dilutable at (ε, n). It is now immediate from (A4) that the degree of dilution is strictly increasing in the equilibrium proportion of output-dependent pay, $\frac{\Delta w^*(N)}{w^*(N)}$. Furthermore, consider a decrease in $p(\varepsilon, n-1)$ due to a decrease in $\alpha(n-1)$ and $\gamma(n-1)$. Since expression (A4) decreases in $\alpha(n-1)$ and $\gamma(n-1)$ (to see the latter, note that $\widehat{p} > \alpha(n-1)$), it follows that the degree of dilution increases when $p(\varepsilon, n-1)$ decreases, holding $p(\varepsilon, n)$ fixed.

We finally prove the additional global result mentioned in footnote 16 in the main text. Assume we can write $\alpha(n-1) = \frac{1}{\kappa_\alpha}\alpha(n)$ and $\gamma(n-1) = \frac{1}{\kappa_\gamma}\gamma(n)$ for all $n \geq 2$ and some $\kappa_\alpha > 1$ and $\kappa_\gamma > 1$. Fixing $\alpha(N)$ and $\gamma(N)$ as before (since $w^*(N)$ and $\Delta w^*(N)$ are taken as exogenously given), it holds that $\alpha(n) = \frac{\alpha(N)}{\kappa_\alpha^{N-n}}$ and $\gamma(n) = \frac{\gamma(N)}{\kappa_\gamma^{N-n}}$ for all $n < N$. Substituting for $\alpha(n)$, $\gamma(n)$ and $\alpha(n-1)$, $\gamma(n-1)$ in (A4) and taking derivatives with respect to κ_α or κ_γ, respectively, shows that dilution increases in κ_α and κ_γ for all $(\varepsilon, n \geq 2)$. □

PROOF OF COROLLARY 2: From expressions (9) and (10), it holds that

$$\Delta w^*(n) - \Delta w^*(n-1) = \frac{\gamma(N)\Delta w^*(N)}{\gamma(n-1)\gamma(n)}\big(\gamma(n-1) - \gamma(n)\big)$$

$$w^*(n) - w^*(n-1) = \left[\frac{\alpha(n-1)}{\alpha(n)} - \frac{\gamma(n-1)}{\gamma(n)}\right]\frac{\alpha(n)\gamma(N)}{\gamma(n-1)}\Delta w^*(N).$$

Hence, $\Delta w\,(n)$ decreases in n if $\gamma(n) > \gamma(n-1)$, and it increases in n if $\gamma(n) < \gamma(n-1)$. Furthermore, $w\,(n)$ decreases in n if $\frac{\gamma(n-1)}{\gamma(n)} > \frac{\alpha(n-1)}{\alpha(n)}$, which defines a cutoff for each n. Note that this cutoff is only relevant for the case of complements, that is, $\gamma\,(n) > \gamma\,(n-1)$. For the case of substitutes, $\gamma\,(n) < \gamma\,(n-1)$, we must have that $\alpha(n) > \alpha\,(n-1)$ for $p\,(\varepsilon, n)$ to increase in n, implying that $\frac{\gamma(n-1)}{\gamma(n)} > \frac{\alpha(n-1)}{\alpha(n)}$ always holds. □

PROOF OF PROPOSITION 4: Since we consider an optimal interior solution, limited liability and monotonicity-in-cash-flow constraints are slack at $\left(w^*(N), \Delta w^*(N)\right)$. Hence, as the relaxed problem minimizes worker rents, expected on-the-job pay $W\,(\varepsilon, N)$ under such a contract will either be tangent to workers' outside option $\underline{w}\,(\varepsilon)$ at some (or many) interior $\varepsilon_N \in \left(\underline{\varepsilon}, \overline{\varepsilon}\right)$ or coincide with it at the two corners $\left\{\underline{\varepsilon}, \overline{\varepsilon}\right\}$.[36] In the former case, it will hold at any point of tangency ε_N that $\frac{\partial}{\partial\varepsilon}\underline{w}(\varepsilon) = \frac{\partial}{\partial\varepsilon}p(\varepsilon, N)\Delta w^*\,(N)$. Hence, $\Delta w^*\,(N) = \frac{\partial}{\partial\varepsilon}\underline{w}(\varepsilon)/\frac{\partial}{\partial\varepsilon}p(\varepsilon, N)\big|_{\varepsilon=\varepsilon_N}$. In the latter case, we must have that $\underline{w}(\overline{\varepsilon}) - \underline{w}(\underline{\varepsilon}) = \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \frac{\partial}{\partial\varepsilon}p(\varepsilon, n)d\varepsilon\Delta w^*\,(N)$. Hence, $\Delta w^*\,(N) = \left(\underline{w}(\overline{\varepsilon}) - \underline{w}(\underline{\varepsilon})\right) / \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \frac{\partial}{\partial\varepsilon}p(\varepsilon, n)d\varepsilon$. In either case, it holds that $w^*\,(N) = \underline{w}\,(\varepsilon_N) - p\,(\varepsilon_N, N)\,\Delta w^*\,(n)$. □

PROOF OF PROPOSITION 5: By optimality for the firm, it must be that each worker's participation constraint binds for at least one realization of ε. Let ε_i be this point for worker i, from which we can express $\Delta w_i = \frac{\underline{w}(\varepsilon_i)-w_i}{p(\varepsilon_i,i)}$. Now, consider two workers, j and j', where $j' > j$. Below we show the claim that it cannot be the case that worker j' receives output-independent pay $w_{j'} > 0$ if worker j receives output-independent pay $w_j < \max_\varepsilon \underline{w}\,(\varepsilon)$—that is, output-independent pay that by itself would be insufficient to satisfy worker j's participation constraint. This claim implies that the scenarios that can arise are that either all workers are paid with output-independent pay only, all workers are paid with output-dependent pay only, or $w_1 = \min\left\{x, \max_\varepsilon \underline{w}\,(\varepsilon)\right\}$ and $w_i = \min\left\{x - \sum_{k=1}^{i-1} w_i, \max_\varepsilon \underline{w}\,(\varepsilon)\right\}$. From these scenarios, it follows that the ratio of output-independent to output-dependent pay must be at least weakly decreasing in i, proving the claim in the proposition.

It remains to show the claim we referred to that if $w_{j'} > 0$, it must be the case that $w_j = \max_\varepsilon \underline{w}\,(\varepsilon)$ for any $j' > j$. Suppose to a contradiction that $w_j < \max_\varepsilon \underline{w}\,(\varepsilon)$ but $w_{j'} > 0$. Consider an alternative contract $(\widetilde{w}_i, \Delta\widetilde{w}_i)_{i=1}^N$ that is the same as $(w_i, \Delta w_i)_{i=1}^N$ for all workers except workers j and j', to which the firm offers $\left(\widetilde{w}_j, \Delta\widetilde{w}_j\right) = \left(w_j + \delta, \frac{\underline{w}(\varepsilon_j)-w_j-\delta}{p(\varepsilon_j,j)}\right)$ and $\left(\widetilde{w}_{j'}, \Delta\widetilde{w}_{j'}\right) = \left(w_{j'} - \delta, \frac{\underline{w}(\varepsilon_{j'})-w_{j'}+\delta}{p(\varepsilon_{j'},j')}\right)$. If δ is sufficiently small, modifying output-independent pay in this manner is feasible because $w_{j'} > 0$ (so lowering $w_{j'}$ is feasible). Furthermore, modifying output-dependent pay in this manner is also feasible because $w_j < \max_\varepsilon \underline{w}\,(\varepsilon)$

[36] The argument for combinations of these two cases follows immediately.

implies that $\Delta w_j > 0$ (so lowering Δw_j is feasible). By construction and feasibility of the original contract, it further holds that $\sum_{i=1}^N \widetilde{w}_i = \sum_{i=1}^N w_i \leq x$ and

$$\sum_{i=1}^{N} \Delta\widetilde{w}_i = \sum_{i=1}^{N} \Delta w_i - \frac{\delta}{p(\varepsilon_j, j)} + \frac{\delta}{p(\varepsilon_{j'}, j')} < \sum_{i=1}^{N} \Delta w_i \leq \Delta x,$$

where the strict inequality follows from $p(\varepsilon_{j'}, j') > p(\varepsilon_j, j)$ since $j' > j$ and $\varepsilon_j = \varepsilon_{j'}$ (due to the imposed regularity conditions).

Under the new contract $(\widetilde{w}_i, \Delta\widetilde{w}_i)_{i=1}^N$, the participation constraints of workers j and j' remain binding at the same points as under the original contract $(w_i, \Delta w_i)_{i=1}^N$. Since these points are either all $\underline{\varepsilon}$ or $\overline{\varepsilon}$ and (given the imposed regularity conditions), we have that $\frac{\partial}{\partial\varepsilon}\left(W_i(\varepsilon, i) - \underline{w}(\varepsilon)\right) > 0$ or $\frac{\partial}{\partial\varepsilon}\left(W_i(\varepsilon, i) - \underline{w}(\varepsilon)\right) < 0$ for any feasible contract, it follows that the participation constraints are strictly satisfied for all $\varepsilon \neq \varepsilon_j = \varepsilon_{j'}$. Hence, we can choose δ sufficiently small such that all participation and feasibility constraints remain satisfied for all ε. It is now immediate from $\sum_{i=1}^N \widetilde{w}_i = \sum_{i=1}^N w_i$ and $\sum_{i=1}^N \Delta\widetilde{w}_i < \sum_{i=1}^N \Delta w_i$ that the firm's aggregate expected compensation cost from offering $(\widetilde{w}_i, \Delta\widetilde{w}_i)_{i=1}^N$ is lower than with $(w_i, \Delta w_i)_{i=1}^N$, providing the desired contradiction. □

PROOF OF PROPOSITION 6: Since all workers $i < N$ must be paid a fixed wage, the firm sets $\Delta w_i = 0$ for all these workers and chooses w_i such that their participation constraint is satisfied. The cheapest such contract for the firm is $(w_i, \Delta w_i)_{i=1}^{N-1} = \left(\max_\varepsilon \underline{w}(\varepsilon), 0\right)$. Since the firm's aggregate compensation costs must be the same as in the relaxed problem, that is, it must be the case that $w^*(N) = \frac{1}{N}\sum_{i=1}^N w_i$ and $\Delta w^*(N) = \frac{1}{N}\sum_{i=1}^N \Delta w_i$, we obtain that the contract offered to worker N must be as stated in the proposition. The necessary and sufficient conditions correspond to the remaining feasibility condition that $w_N \geq 0$ (which is also sufficient for the feasibility of the contracts offered to $i < N$) and the participation constraint of worker N,

$$Nw^*(N) - (N-1)\max_{\varepsilon} \underline{w}(\varepsilon) + p(\varepsilon, N)N\Delta w^*(N) \geq \underline{w}(\varepsilon).$$

□

PROOF OF LEMMA 3:

(i) We show that absent feasibility constraints, asymmetric contracts that do not condition on n always entail costs for the firm. In the main text, we have already shown the claim for the case of idiosyncratic shocks $\frac{\partial\underline{w}(\varepsilon)}{\partial\varepsilon} = 0$. [37] We now extend the claim to the case of systematic shocks with $\frac{\partial\underline{w}(\varepsilon)}{\partial\varepsilon} > 0$. Specifically, take an optimal contract of the form

[37] We obtain a similar prediction if $\frac{\partial\underline{w}(\varepsilon)}{\partial\varepsilon} < 0$, but note that in this case, the limited liability constraint is always binding.

$(w_i, \Delta w_i)_{i=1}^N$ that does not condition on n and for which staying is an iteratively dominant strategy for each worker. For every i, let ε_i be the shock realization for which $w_i + p(\varepsilon, n)\Delta w_i - \underline{w}(\varepsilon)$ is minimal. Let

$$\Delta y_i = \frac{p(\varepsilon_i, i)}{p(\varepsilon_i, N)}\Delta w_i \leq \Delta w_i,$$

where the inequality is strict if $\Delta w_i > 0$ and $i \neq N$. By construction, it holds that $w_i + p(\varepsilon, N)\Delta y_i = w_i + p(\varepsilon, i)\Delta w_i \geq \underline{w}(\varepsilon)$ for every (ε, n). Now, construct a symmetric contract that conditions on n for which $(\widetilde{w}(N), \Delta\widetilde{w}(N)) = \left(\sum_{i=1}^N \frac{w_i}{N}, \sum_{i=1}^N \frac{\Delta y_i}{N}\right)$. By construction, it holds that the firm's compensation cots at full retention are

$$\sum_{i=1}^N \mathrm{E}(w_i + p(\varepsilon, N)\Delta w_i) \geq \sum_{i=1}^N \mathrm{E}(w_i + p(\varepsilon, N)\Delta y_i) \quad \text{(A5)}$$
$$= N\mathrm{E}(\widetilde{w}(N) + p(\varepsilon, N)\Delta\widetilde{w}(N)),$$

where E denotes the expectation with respect to the shock realization ε. The inequality in (A5) is strict if $\Delta w_i > 0$ for at least one $i < N$. Since $\frac{\partial \underline{w}(\varepsilon)}{\partial \varepsilon} > 0$, this must indeed be the case for any contract matching the compensation costs from the relaxed problem (Proposition 4). Thus, the firm's aggregate compensation costs under the symmetric contract are strictly lower. Constructing all out-of-equilibrium promises $(\widetilde{w}(n), \Delta\widetilde{w}(n))_{n=1}^{N-1}$ as dictated by Proposition 3, the claim follows.

(ii) In what follows, we consider the case of $x = 0$ and provide a concrete condition for when the costs of symmetric contracts are sufficiently higher than those of asymmetric contracts that do not condition on n. Since $w(n) \leq \frac{x}{n} = 0$, we obtain that $w(n) = 0$ for all n. From Lemma I.A1 in the Internet Appendix, it follows that $W(\varepsilon, n+1) = W(\varepsilon, n)$ for all n for at least one $\varepsilon_n \in [\underline{\varepsilon}, \overline{\varepsilon}]$. Hence, $\Delta w(n+1)$ is pinned down as $\Delta w(n+1) = \frac{p(\varepsilon_n, n)}{p(\varepsilon_{n+1}, n+1)}\Delta w(n)$. By construction, $\Delta w(n+1) \geq 0$. Furthermore,

$$\Delta x - \frac{p(\varepsilon_n, n)}{p(\varepsilon_{n+1}, n+1)}\Delta w(n) \geq \frac{p(\varepsilon_{n+1}, n+1)\Delta x - p(\varepsilon_n, n)\Delta x}{p(\varepsilon_{n+1}, n+1)} \geq 0,$$

where the last inequality follows from the assumption that retaining workers is efficient for the firm. Hence, $\Delta w(n+1)$ is feasible if $\Delta w(n)$ is feasible. Now, note that

$$\frac{\partial}{\partial \varepsilon}\big(W(\varepsilon, n+1) - W(\varepsilon, n)\big) = \left(p(\varepsilon, n+1)\frac{p(\varepsilon_n, n)}{p(\varepsilon_n, n+1)} - p(\varepsilon, n)\right)\Delta w(n),$$

which is positive if and only if $\frac{\partial}{\partial \varepsilon}\frac{p(\varepsilon, n+1)}{p(\varepsilon, n)} = \left(\frac{\gamma(n+1)}{\gamma(n)} - \frac{\alpha(n+1)}{\alpha(n)}\right)\frac{\gamma(n)\alpha(n)\frac{\partial}{\partial \varepsilon}\beta(\varepsilon)}{p(\varepsilon, n)^2} \geq 0$. It follows that $\varepsilon_n = \underline{\varepsilon}$ if $\frac{\gamma(n+1)}{\gamma(n)} \geq \frac{\alpha(n+1)}{\alpha(n)}$, and $\varepsilon_n = \overline{\varepsilon}$ if $\frac{\gamma(n+1)}{\gamma(n)} < \frac{\alpha(n+1)}{\alpha(n)}$.

From this, we immediately obtain

$$\Delta w(n) = \prod_{i=1}^{n-1} \frac{p(\underline{\varepsilon}, i)}{p(\underline{\varepsilon}, i+1)} \mathbf{1}_{\underline{\varepsilon}} \frac{p(\overline{\varepsilon}, i)}{p(\overline{\varepsilon}, i+1)} (1 - \mathbf{1}_{\underline{\varepsilon}}) \Delta w(1), \tag{A6}$$

where we define the indicator function $\mathbf{1}_{\underline{\varepsilon}} = 1$ if $\frac{\gamma(n+1)}{\gamma(n)} \geq \frac{\alpha(n+1)}{\alpha(n)}$ and $\mathbf{1}_{\underline{\varepsilon}} = 0$ otherwise. Note that $\Delta w\,(n)$ in expression (A6) is decreasing in n, since $\frac{p(\varepsilon,i)}{p(\varepsilon,i+1)} < 1$. Furthermore, since $w\,(n) = 0$ and $\Delta w\,(n)$ decreases in n, we obtain that any optimal contract is dilutable for all (ε, n). Finally, from the binding participation constraint for $n = 1$, we obtain $\Delta w\,(1) = \frac{\underline{w}}{p(\varepsilon_1,1)}$, where ε_1 is the highest value for ε for which the resulting expected compensation $W(\varepsilon, 1)$ touches but does not cross $\underline{w}(\varepsilon)$ on the interval $[\underline{\varepsilon}, \overline{\varepsilon}]$. For example, in the case of idiosyncratic risk, $\varepsilon_1 = \underline{\varepsilon}$.

It remains to compare the optimal symmetric contract defined above with the optimal asymmetric contract that does not condition on n. To construct the latter contract, assign a number from 1 to N to each worker (since the workers are identical, the order is arbitrary) and offer each worker $n \in \{1, \ldots, N\}$ the asymmetric contract $(w_n, \Delta w_n) = \left(0, \frac{\underline{w}(\widehat{\varepsilon}_n)}{p(\widehat{\varepsilon}_n, n)}\right)$, where $\widehat{\varepsilon}_n$ is the highest value of ε for which the resulting expected compensation $p\,(\varepsilon, n)\,\frac{\underline{w}(\widehat{\varepsilon}_n)}{p(\widehat{\varepsilon}_n, n)}$ touches but does not cross $\underline{w}(\varepsilon)$ on the interval $[\underline{\varepsilon}, \overline{\varepsilon}]$. For example, in the case of idiosyncratic risk, $\widehat{\varepsilon}_n = \underline{\varepsilon}$ for all n. Note that such an asymmetric contract makes staying an iteratively dominant strategy for workers at the lowest cost for the firm. This contract leads to a lower expected compensation cost than the optimal symmetric contract if and only if

$$\frac{1}{N}\left(\sum_{n=1}^{N} \frac{\underline{w}(\widehat{\varepsilon}_n)}{p(\widehat{\varepsilon}_n, n)}\right) < \prod_{m=1}^{N-1} \frac{p(\underline{\varepsilon}, m)}{p(\underline{\varepsilon}, m+1)} \mathbf{1}_{\underline{\varepsilon}} \frac{p(\overline{\varepsilon}, m)}{p(\overline{\varepsilon}, m+1)} (1 - \mathbf{1}_{\underline{\varepsilon}}) \frac{\underline{w}(\varepsilon_1)}{p(\varepsilon_1, n)}. \tag{A7}$$

□

PROOF OF PROPOSITION 7: Consider offering the optimal symmetric contract to $N - 1$ workers and the optimal contract from the relaxed problem to the final worker. Doing so is feasible, as the original and new contracts offered to each worker respect the feasibility constraints $0 \leq w_i\,(n) \leq \frac{x}{n}$ and $0 \leq \Delta w_i\,(n) \leq \frac{\Delta x}{n}$ and are monotonic-in-retention. Furthermore, if the optimal symmetric contract is associated with positive coordination costs, it holds that $W\,(\varepsilon, N) > W^*\,(\varepsilon, N)$, implying that the average per-worker cost drops to $\frac{1}{N}\,((N-1)\,W\,(\varepsilon, N) + W^*\,(\varepsilon, N)) < W\,(\varepsilon, N)$. This argument completes the proof.

For interested readers, we discuss the construction of the optimal asymmetric full retention contract in the Internet Appendix.[38] □

[38] Under realistic conditions, the firm offers the optimal symmetric full-retention contract to worker $i = 1$, the optimal contract from the relaxed problem to worker $i = N$, and all other

REFERENCES

Aldatmaz, Serdar, Paige Ouimet, and Edward D. Van Wesep, 2018, The option to quit: The effect of employee stock options on turnover, *Journal of Financial Economics* 127, 136–151.

Barnette, Thomas, Mary Fink, Greg Freeze, William Mahan, Danny Nelms, Kim Nowell, and Luke Viera, 2024, *2024 Retention Report: Decoding the Emerging Workforce to Accelerate Retention, Engagement, and Profits* (Work Institute, Franklin, Tennessee).

Bergman, Nittai K., and Dirk Jenter, 2007, Employee sentiment and stock option compensation, *Journal of Financial Economics* 84, 667–712.

Berk, Jonathan B., Richard Stanton, and Josef Zechner, 2010, Human capital, bankruptcy, and capital structure, *Journal of Finance* 65, 891–926.

Bettis, J. Carr, John Bizjak, Jeffrey L. Coles, and Swaminathan Kalpathy, 2018, Performance-vesting provisions in executive compensation, *Journal of Accounting and Economics* 66, 194–221.

Breitling, Frank, Julia Dhar, Ruth Ebeling, and Deborah Lovich, 2021, 6 Strategies to Boost Retention Through the Great Resignation, *Harvard Business Review*, November 15.

Chen, Alvin, 2024, Firm performance pay as insurance against promotion risk, *Journal of Finance* 79, 3497–3541.

Corsetti, Giancarlo, Amil Dasgupta, Stephen Morris, and Hyun Song Shin, 2004, Does one Soros make a difference? A theory of currency crises with large and small traders, *Review of Economic Studies* 71, 87–113.

Derler, Andrea, Ian Cook, Mike Everitt, Carlina Kim, Macguire Rintoul, and Anton Smessaert, 2023, Turnover contagion is real, Visier Insights report.

Diamond, Douglas W., and Philip H. Dybvig, 1983, Bank runs, deposit insurance, and liquidity, *Journal of Political Economy* 91, 401–419.

Döttling, Robin, Tomislav Ladika, and Enrico Perotti, 2019, Creating intangible capital, Working paper, University of Amsterdam.

Edmans, Alex, Xavier Gabaix, and Dirk Jenter, 2017, Executive compensation: A survey of theory and evidence, in Benjamin Hermalin and Michael Weisbach, eds.: *The Handbook of the Economics of Corporate Governance* (Elsevier, Amsterdam, the Netherlands).

Felps, Will, Terence R. Mitchell, David R. Hekman, Thomas W. Lee, Brooks C. Holtom, and Wendy S. Harman, 2009, Turnover contagion: How coworkers' job embeddedness and job search behaviors influence quitting, *Academy of Management Journal* 52, 545–561.

Fulghieri, Paolo, and David Dicks, 2024, Uncertainty, contracting, and beliefs in organizations, *Review of Financial Studies*, forthcoming.

Erickson, Robin, Alice Kwan, Neil Neveras, Bill Pelster, Jeff Schwartz, and Sara Szpaichler, Talent, 2020, Surveying the talent paradox from the employee perspective, Deloitte.

Gittleman, Maury, 2022, *The "Great Resignation" in perspective, Monthly Labor Review*, U.S. Bureau of Labor Statistics, July.

Goldstein, Itay, and Ady Pauzner, 2005, Demand-deposit contracts and the probability of bank runs, *Journal of Finance* 60, 1293–1327.

Halac, Marina, Ilan Kremer, and Eyal Winter, 2020, Raising capital from heterogeneous investors, *American Economic Review* 110, 889–921.

Halac, Marina, Elliot Lipnowski, and Daniel Rappoport, 2021, Rank uncertainty in organizations, *American Economic Review* 111, 757–786.

Hancock, Jullie I., David G. Allen, Frank A. Bosco, Karen R. McDaniel, and Charles A. Pierce, 2013, Meta-analytic review of employee turnover as a predictor of firm performance, *Journal of Management* 39, 573–603.

Hand, John R. M., 2008, Give everyone a prize? Employee stock options in private venture-backed firms, *Journal of Business Venturing* 23, 385–404.

workers receive compensation in equilibrium that is in between the compensation offered to these two workers.

Hausknecht, John P., and Charlie O. Trevor, 2011, Collective turnover at the group, unit, and organizational levels: Evidence, issues, and implications, *Journal of Management* 37, 352–388.

Heavey, Angela L., Jacob A. Holwerda, and John P. Hausknecht, 2013, Causes and consequences of collective turnover: A meta-analytic review, *Journal of Applied Psychology* 98, 412–453.

Hoberg, Gerard, and Gordon Phillips, 2016, Text-based network industries and endogenous product differentiation, *Journal of Political Economy* 124, 1423–1465.

Hochberg, Yael V., and Laura Lindsey, 2010, Incentives, targeting and firm performance: An analysis of non-executive stock options, *Review of Financial Studies* 23, 4148–4186.

Innes, Robert D., 1990, Limited liability and incentive contracting with ex-ante action choices, *Journal of Economic Theory* 52, 45–67.

Jochem, Torsten, Tomislav Ladika, and Zacharias Sautner, 2018, The retention effects of unvested equity: Evidence from accelerated option vesting, *Review of Financial Studies* 31, 4142–4186.

Kedia, Simi, and Shiva Rajgopal, 2009, Neighborhood matters: The impact of location on broad based stock option plans, *Journal of Financial Economics* 92, 109–127.

Lazear, Edward, 2004, Output-based pay: Incentives, retention or sorting?, in Solomon W. Polachek, ed.: *Accounting for Worker Well-Being* (Emerald Group Publishing Limited, Leeds, United Kingdom).

Luo, Dan, and Ming Yang, 2024, The optimal structure of securities under coordination frictions, Working paper, University College London.

McFeely, Shane, and Ben Wigert, 2019, *This Fixable Problem Costs U.S. Businesses $1 Trillion*, Gallup, March.

Morris, Stephen, and Hyun S. Shin, 2003, Global games: Theory and applications, in Mathias Dewatripont, Lars P. Hansen, and Stephen J. Turnovsky, eds.: *Advances in Economics and Econometrics* (Cambridge University Press, Cambridge).

Murphy, Kevin J., 2003, Stock-based pay in new economy firms, *Journal of Accounting and Economics* 34, 129–147.

Oyer, Paul, 2004, Why do firms use incentives that have no incentive effects?, *Journal of Finance* 59, 1619–1649.

Rajan, Raghuram G., 2012, Presidential address: The corporation in finance, *Journal of Finance* 67, 1173–1217.

Sakovics, Jozsef, and Jakub Steiner, 2012, Who matters in coordination problems?, *American Economic Review* 102, 3439–3461.

Terovitis, Spyros, and Vladimir Vladimirov, 2024, How financial markets create superstars, Working paper, University of Amsterdam.

Titman, Sheridan, 1984, The effect of capital structure on a firm's liquidation decision, *Journal of Financial Economics* 13, 137–151.

Winter, Eyal, 2004, Incentives and discrimination, *American Economic Review* 94, 764–773.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1: Internet Appendix.
Replication Code.