DISCUSSION PAPER SERIES

DP17419

Worker Runs

Florian Hoffmann and Vladimir Vladimirov

FINANCIAL ECONOMICS

ISSN 0265-8003

Worker Runs

Florian Hoffmann and Vladimir Vladimirov

Discussion Paper DP17419
Published 30 June 2022
Submitted 30 June 2022

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

Worker Runs

Abstract

The voluntary departure of hard-to-replace skilled workers worsens firm prospects, thus, increasing remaining workers' incentives to leave. We develop a model of collective turnover in which firms design compensation to limit the risk of such "worker runs." To achieve cost-efficient retention, firms may use fixed or dilutable variable pay -- such as stock option/bonus pools -- that promises remaining workers more when others leave but gets diluted otherwise. The optimal mix of fixed and dilutable pay depends on firms' relative risk exposure and their financial constraints. Compensating (identical) workers differently and financing investments with debt can improve collective retention.

JEL Classification: G32, M52, J54, J33

Keywords: compensation structure of non-executive employees, high-skilled employees, contagious turnover, worker runs, worker bargaining power, financing labor

Florian Hoffmann - florian.hoffmann@kuleuven.be
KU Leuven

Vladimir Vladimirov - vladimirov@uva.nl
University of Amsterdam and CEPR

Worker Runs*

Florian Hoffmann[†] Vladimir Vladimirov[‡]

June 30, 2022

Abstract

The voluntary departure of hard-to-replace skilled workers worsens firm prospects, thus, increasing remaining workers' incentives to leave. We develop a model of collective turnover in which firms design compensation to limit the risk of such "worker runs." To achieve cost-efficient retention, firms may use fixed or dilutable variable pay – such as stock option/bonus pools – that promises remaining workers more when others leave but gets diluted otherwise. The optimal mix of fixed and dilutable pay depends on firms' relative risk exposure and their financial constraints. Compensating (identical) workers differently and financing investments with debt can improve collective retention.

Keywords: Compensation structure of non-executive employees, high-skilled employees, contagious turnover, worker runs, worker bargaining power.

JEL Classification: G32, M52, J54, J33

*We thank Dan Bernhardt, Philip Bond, Arnoud Boot, Matthieu Bouvard, Tolga Caskurlu, Alvin Chen, Paolo Fulghieri, Uli Hege, Roman Inderst, Ben Iverson, Torsten Jochem, Tomislav Ladika, Tingjun Liu, Andrey Malenko, Rich Mathews (Cambridge discussant), Yuliyan Mitkov, Stefan Obernberger, Martin Oehmke, Enrico Perotti, Konrad Raff (EFA discussant), Tano Santos, Elena Simintzi, Martin Szydlowski, Luke Taylor, 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 (Northwestern), Toulouse School of Economics, University of Amsterdam, University of Bristol, and University of Warwick for constructive feedback. An old version of the paper that contained some of the results circulated under the title "Financing Skilled Labor."

[†]KU Leuven, email: florian.hoffmann@kuleuven.be.

[‡]University of Amsterdam and CEPR, e-mail: vladimirov@uva.nl.

1 Introduction

In June 2021, the financial press extensively reported on Credit Suisse's fight to stem the exodus of senior and junior employees across multiple divisions. The trigger for this exodus was the bank's exposure to the spectacular collapses of Archegos and Greensill Capital, which dented its profits. Senior bankers 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 started leaving in droves, as "nobody wanted to be the last man standing." Making matters worse, Credit Suisse's stock had plunged by 16% while that of its main rivals had risen, "dealing a further blow to staff who have seen the value of their deferred pay in stock diminish."[1]

The retention problems faced by Credit Suisse illustrate a broader phenomenon. Over the last twenty years, 23% of workers have voluntarily quit their jobs every year, with the recent spike in voluntary turnover dubbed the "Great Resignation" by the popular press (Breitling et al., 2021). The spike is highest in human-capital-intensive industries, ranging from hospitality to high-tech and professional business services (BLS, 2021). While voluntary quits can be good from an efficiency perspective if they lead to a better match between firms and workers, they can also be detrimental when they are contagious, i.e., the result of a coordination failure among workers in fundamentally sound firms. Evidence suggests that voluntary turnover often happens in waves, with multiple workers leaving within a short time period (Felps et al., 2009; Hausknecht and Trevor, 2011; Heavey et al., 2013; Hancock et al., 2013). The cost to firms is staggering. In the U.S. alone, the annual cost related to replacement, training, and lost productivity is reported at over a trillion dollars (Gallup, 2019; Work Institute, 2019). Thus, a better understanding of how to reduce this cost by preventing inefficient turnover is key for firms. Credit Suisse estimates that reducing turnover can save it $100 million per year (Graves, 2016).

In this paper, we develop a model of collective turnover and characterize how compensation design and firm financing can help firms retain (groups of) skilled employees. The main novelty of our analysis lies in capturing strategic complementarities in workers' decisions to leave, i.e., the problem that the departure of some workers (or the risk of such departures) makes other workers more likely to leave as well, leading workers to leave even fundamentally healthy firms. Such "worker runs" are both particularly likely and particularly costly in firms that rely heavily on teams of hard-to-replace workers with complementary skills. Typical examples are startups, whose success crucially depends on retaining well-functioning teams consisting of key personnel in R&D, management, sales and marketing. Further examples

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

include consulting, advisory, investment banking, law firms, and PE partnerships.

In our model, the fundamental reason workers may want to leave is that, after being hired, they learn about the firm's prospects by privately observing signals about shocks to its productivity. These signals affect workers' beliefs about how their on-the-job pay compares to available outside options and may trigger some workers to leave. Crucially, the departure of skilled employees makes other workers also more likely to leave. The reason for these strategic complementarities in workers' decisions to leave is that any loss of hard-to-replace human capital dents firm productivity and, thus, the value of remaining workers' (variable) compensation.

In line with the Credit Suisse example, we focus our analysis on inefficient collective turnover in the form of a "worker run" that the firm wants to prevent. In particular, we investigate how the risk of such a run can be mitigated via optimally designed compensation, i.e., via contracts that achieve any given retention level at the lowest compensation costs. Our focus on this compensation design problem is motivated by evidence that compensation is both a primary reason for voluntary quits and the most important retention tool (Payscale, 2019; Deloitte, 2020; Breitling et al., 2021). Three model ingredients drive our main compensation design predictions: First, productivity shocks are not observable, implying that workers earn rent on their private signals about how their on-the-job pay compares to their outside options. Second, compensation design needs to account for the strategic complementarities in workers' decisions to leave. Third, compensation design is restricted by constraints on the resources available for compensating the firm's workforce. The importance of these constraints is suggested by the fact that the median ratio of compensation expenses to capital expenditures for Compustat firms in 2021 is 9.8.

Mitigating the strategic complementarities in workers' decisions to stay or leave requires smoothing workers' expected compensation across different retention scenarios. Deferred fixed compensation that pays the same in all cash flow states is particularly suitable, as its value is independent of the retention of other workers.[2] This is not the case for compensation contingent on firm performance, such as performance bonuses or equity-based pay. To counteract that the value of such compensation decreases with the loss in human capital associated with worker departures, the firm can optimally design compensation in a way that promises remaining workers higher pay when other workers are leaving. Examples of such "dilutable" compensation include paying workers from a stock (option) pool or offering profit-sharing bonuses.[3] Since workers that leave forgo their deferred compensation, remain-

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

[3] In practice, the size of such bonus pools is usually a percentage of the firm's net revenues and fluctuates with cash flows, akin to equity compensation. Indeed, bonuses can be shared in the form of either stocks or a cash amount. Hence, for brevity, we refer to such compensation as "equity-based."

ing workers mechanically incur less dilution on their claims, which reduces their incentives to leave. Conversely, if the firm is able to retain more workers, thereby increasing its probability of success, the workers' promised variable pay can be diluted without endangering retention, as the probability of actually obtaining such compensation increases.

The optimal mix of fixed and dilutable (equity-based) compensation depends on their relative cost to the firm, which in turn is determined by the firm's sensitivity to productivity shocks relative to that of workers' outside options. To see this, consider the case in which shocks affect the firm's prospects much more than workers' outside options. Idiosyncratic shocks are the prime example. In this case, the firm faces two challenges. First, it needs to improve worker retention after negative shocks that reduce the value of workers' variable pay relative to their outside options. Second, the firm wants to avoid overpaying workers following positive shocks, which increase workers' on-the-job pay relative to their outside options. The best way of achieving both objectives is to offer fixed compensation. Fixed compensation improves retention by insulating workers both from the departure of other workers and from shocks that affect the firm's growth prospects. As a result, compensation matches workers' outside options as close as possible, which insures retention at the lowest possible cost.

Relying purely on fixed compensation ceases to be optimal in the case of common shocks that also affect the value of workers' outside options. Indeed, if workers' outside options are more sensitive to shocks than what the firm has offered, the firm's problem is to retain workers following positive (instead of negative) productivity shocks that improve their outside options more than their on-the-job pay. Hence, the firm needs to *increase* the sensitivity of compensation to systematic shocks by using less fixed and more variable (dilutable) pay. Doing so lowers worker rents without having to sacrifice efficiency. In particular, workers' compensation increases in tandem with their outside options, which ensures retention, and decreases when their outside options are low, which avoids overpaying workers that stay anyway.[4] Taken together, a lower *sensitivity* to (common) risk makes it optimal to rely more on dilutable (e.g., equity-based) and less on fixed pay.

Another key lever through which firms can improve the retention of groups of skilled workers is by offering different compensation contracts to ex-ante identical workers. In practice, this means that even a marginally higher benefit of retaining some workers can lead to large differences in compensation design. Intuitively, contagious turnover can be reduced or

[4] Call options are particularly suitable to boost the sensitivity of on-the-job pay to such systematic shocks. While we do not consider offering equity-based pay for incentive reasons, the effect of such considerations on the optimal fixed-dilutable pay mix is straightforward. However, note that for workers below the executive level, the incentives effects of variable (equity-based) pay may be muted due to free-riding opportunities (Holmström, 1982).

even prevented if workers are convinced that other workers will never leave the firm. This makes targeting a subset of workers with compensation contracts that make them less likely to leave an attractive option. Retaining some workers with higher probability does not necessarily mean they are paid more, as it can also be achieved by structuring compensation differently. Specifically, if the firm is highly exposed to idiosyncratic or systematic shocks, it can improve retention of some workers by making their compensation safer, i.e., less sensitive to both shocks and the retention of other workers. The resulting reduction in (strategic) uncertainty regarding these workers' decision to stay or leave can make it more likely that the remaining workers will be retained.

The firm's choice of financing also affects the cost of improving collective retention. We show that firms with low sensitivity to shocks (relative to workers' outside options) can achieve more cost-efficient retention by raising debt financing. Doing so allows the firm to preserve more of its equity for compensation purposes, which is necessary to match better the relatively high sensitivity of workers' outside options to shocks. By contrast, a firm with high risk exposure is better off raising equity financing. In this case, preserving the firm's limited resources in low cash flow states for paying fixed wages to workers rather than repaying investors allows the firm to make compensation safer and, thus, more aligned with workers' outside options.

Related Literature. While the staggering cost to firms that arise from contagious collective turnover is widely discussed by the popular press and the management literature (Felps et al., 2009; Hausknecht and Trevor, 2011; Hancock et al., 2013; Heavey et al., 2013), to the best of our knowledge, our paper is the first to formally model this problem and to derive implications for a firm's compensation and financing policies. A key problem of collective turnover is that it can be inefficient, as workers do not internalize the impact of their decisions on other workers. We model this idea as a coordination problem similar to the bank-run literature (Diamond and Dybvig, 1983; Goldstein and Pauzner, 2005). Our focus on mitigating runs through compensation design is due to the fact that standard solutions such as deposit insurance, mandatory stay, and suspension of convertibility have no obvious analog in the context of retaining workers.

Our focus on worker runs as a coordination problem that can also affect fundamentally healthy firms (as illustrated by our Credit Suisse example) complements work in which workers leave firms because of concerns about their financial health (Titman, 1984; Berk et al., 2010). The main idea in this literature is that the higher risk of default associated with higher leverage makes retention more costly.[5] By contrast, we show that this prediction

[5] Also related are Döttling et al. (2020), who argue that firms relying on intangible capital should be more financially prudent, as that helps to insure workers' compensation.

reverses for firms with low sensitivity to systematic risk. In that case, using debt financing allows the firm to preserve more equity for compensation purposes and helps it better align workers' on-the-job pay with their outside options. Our paper also relates to Bolton et al. (2019), who study how a firm can improve the retention of a single risk-averse entrepreneur who cannot commit to staying. The main conceptual difference is that Bolton et al. (2019) study compensation contracts that are continuously fine-tuned in complete markets in which firms can effectively contract on shock realizations. In contrast, productivity shocks are not contractible in our model, making such fine-tuning impossible. This delivers a rich set of novel compensation design predictions depending on whether shocks are idiosyncratic or systematic.

Our result that firms might improve retention by compensating identical workers differently expands on papers that analyze the consequences of heterogeneous payoffs in games of strategic complementarity (Corsetti et al., 2004). Similar to that work, we show that heterogeneity in payoffs can mitigate coordination problems and reduce the incidence of runs. However, while prior work assumes that agents' payoff functions are exogenously given, our innovation is that firms can endogenously introduce heterogeneity in payoffs by offering different contracts. Our main contribution then is to characterize how the different contracts offered to different groups of (identical) workers are optimally structured. This focus on how a different compensation *structure* can mitigate contagious turnover also differentiates our paper from Winter (2004) and Halac et al. (2021), who analyze the role of different compensation *levels* in avoiding coordination problems in teams.

By modeling strategic complementarities in workers' decisions to leave as well as resource constraints on the firm's overall wage bill, we also offer several new insights complementing papers that analyze the retention of a single worker. While the idea that firms try to match workers' outside options through a combination of fixed and equity pay is also discussed in Oyer (2004), the trade-offs in his model — in which the firm tries to insure a risk-averse worker — are different. In particular, Oyer (2004) obtains that firms with a higher exposure to systematic risk should use less fixed and more equity compensation. Qualitatively, this prediction is opposite to the one we obtain in our model in which workers are risk-neutral, and optimal compensation design for a given retention level is determined by rent-extraction arguments. Concretely, the difference in predictions is driven by the fact that compensation in Oyer (2004) is restricted always to be less sensitive to systematic shocks than the worker's outside options in order to insure the risk-averse worker.[6] However, even if were to impose

[6] In particular, Oyer (2004) shows that offering insurance to a risk-averse worker becomes more expensive as the firm's sensitivity to common shocks increases. Hence, firms offer *more* equity as their sensitivity to common shocks increases. By contrast, we show that when firms become more sensitive to common shocks, matching workers' outside options and, thus, reducing rents requires *less* equity-based pay.

risk-aversion in our model, such a restriction may not be technologically feasible or may violate the firm's resource constraints in our multiple worker setting.

Finally, our analysis highlights the importance of dilution — a fundamental feature of, e.g., equity-based pay that is often neglected in theoretical work. In particular, we show that dilutable equity-based pay can help *lower* 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 suitable for such firms despite arguments to the contrary in prior work that does not account for the dilution benefits of equity pay (Murphy, 2003; Lazear, 2004; Oyer, 2004).[7]

2 Baseline Model

We develop a tractable compensation design model of a firm seeking to retain a group of N workers indexed by $i \in \{1, ..., N\}$. All parties are risk-neutral and do not discount future payoffs.

Project. The firm has liquid assets in place $x > 0$ and a project that requires hiring $N \geq 2$ skilled workers at $t = 0$ in order 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 on two values: $\Delta x > 0$ corresponding to project success or zero corresponding to failure.[8] The probability of success $p = p(\varepsilon, n) \in (0, 1)$ depends on the realization of an exogenous (productivity) shock ε realized at the interim date, $t = 1$, and on the number of workers $0 \leq n \leq N$ that stay with the firm until $t = 2$, i.e., that do not leave at $t = 1$. To capture the impact of human capital on project success, we stipulate that p is increasing in n, i.e., $p(\varepsilon, n-1) \leq p(\varepsilon, n)$ for $1 < n < N$ and all ε. 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, $\frac{\partial}{\partial \varepsilon} p(\varepsilon, n) \geq 0$, capturing either

[7] More broadly, our results add a novel perspective to work studying why firms offer equity-based compensation to workers below the executive level, given that such compensation is unlikely to have any incentive effects (Holmström, 1982). While deferring pay to improve retention is seen as uncontroversial, the question of what type of pay to defer is less clear-cut. Prior theory has explained the use of (deferred) equity-based compensation for the purpose of: avoiding wage renegotiations when the firm's equity value is correlated with the workers' outside options (Oyer, 2004); aligning the incentives of managers with the interests of investors (Lazear, 2004); exploit the overoptimism of boundedly rational workers (Bergman and Jenter, 2007); providing a hedge against not being promoted (Chen, 2020); or hedging uncertainty under Knightian preferences (Fulghieri and Dicks, 2019).

[8] An equivalent specification is to set assets in place to zero and assume that the project generates cash flows of x in case of failure and of $x + \Delta x$ in case of success.

positive firm-specific developments or favorable market-wide conditions.

Interim information and workers' decision. The true realization of the productivity shock is not observable to anyone, but for any given shock realization, ε, each worker i privately observes a signal $\widetilde{\varepsilon}_i = \varepsilon + \sigma e_i$ at $t = 1$, where σ is small but positive, i.e., $\sigma > 0$, $\sigma \to 0$, and the zero mean noise terms e_i are iid draws from a distribution F that admits a density f with support on the real line.[9] As is standard, our assumption is that the induced distribution over the shock realization ε conditional on the observed signal $\widetilde{\varepsilon}_i$ satisfies the monotone likelihood ratio property, i.e., for any $\widetilde{\varepsilon}_i'' > \widetilde{\varepsilon}_i'$ and $\varepsilon'' > \varepsilon'$, it holds that $\frac{g\left(\varepsilon''|\widetilde{\varepsilon}_i''\right)}{g\left(\varepsilon''|\widetilde{\varepsilon}_i'\right)} > \frac{g\left(\varepsilon'|\widetilde{\varepsilon}_i''\right)}{g\left(\varepsilon'|\widetilde{\varepsilon}_i'\right)}$, such that a higher signal is indicative of a higher realization of ε. Upon observing the private signal, each worker decides whether to stay at the firm or take an outside option of value $\underline{w}(\varepsilon)$. One may think of this outside option as the expected value to a worker from starting an own business or joining another firm. Without loss of generality, we assume that in case of indifference, a worker prefers to stay with the firm at which she is currently employed, which could be rationalized with an arbitrary small switching cost. In our baseline specification, we consider the case in which workers take the decision to stay or leave the firm simultaneously at $t = 1$. We discuss sequential decisions as an extension in Section 4.1.2.

To organize the analysis going forward, it will be useful to distinguish between two cases based on how the exogenous shock ε affects value creation within and outside the firm. In particular, we say that the shock ε captures *idiosyncratic risk* if it only affects value creation within the firm — i.e., $p(\varepsilon, n)$ 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 the workers' outside options, in which case we stipulate that $\underline{w}(\varepsilon)$ is monotonically increasing $\frac{\partial}{\partial\varepsilon}\underline{w}(\varepsilon) \geq 0$ with a finite limit.[10]

We focus our main analysis on the problem of *efficient retention* by making the following three assumptions: First, the number of workers the firm is hiring, N, is given. Second, we stipulate that higher retention (higher n) is efficient in the sense of maximizing the expected joint surplus of the firm and its (retained) workers. That is, for all ε, the net surplus from retaining n workers at $t = 1$,

$$\Omega(\varepsilon, n) := p(\varepsilon, n)\,\Delta x - n\underline{w}(\varepsilon), \tag{1}$$

is positive and increasing in n, i.e., $\Omega(\varepsilon, n) \geq \Omega(\varepsilon, n-1)$ for $1 \leq n \leq N$.[11] The implicit as-

[9] Whether the firm's owners also observe a private signal about ε is inconsequential for our main analysis.

[10] For analytical convenience, we do not consider idiosyncratic and systematic shocks simultaneously. Doing so would not add any further qualitative insights. Ultimately, what matters for most of our subsequent results is whether productivity shocks affect the firm more or less than workers' outside options.

[11] 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 by making contracts easy to terminate. This is

sumption is that the firm cannot replace workers who leave at $t = 1$ with equally productive ones. This assumption captures the problem that finding, hiring, and training replacements for skilled workers is a costly and time-consuming process, which dents the firm's productivity. Finally, we assume that workers can at most realize $\mathrm{E}[w(\varepsilon)]$ as their $t = 0$ outside option, which implies that their ex ante participation constraints are never binding. We discuss the implications of relaxing these assumptions further below.

Compensation contracts. Compensation contracts are signed at $t = 0$. Workers who leave at $t = 1$ forgo all their compensation, i.e., compensation is deferred, which is always optimal in our setting, as it helps retention and does not make hiring more difficult. When staying with the firm until $t = 2$, each worker is paid according to a compensation contract $C := \{w(n), \Delta w(n)\}_{n=1}^{N}$. This contract stipulates a transfer of $w(n)$ to the worker in the low cash flow state and $w(n) + \Delta w(n)$ in the high cash flow state, depending on the number of workers n the firm can retain.[12] The main contracting friction is that neither the productivity shock ε nor signals $\widetilde{\varepsilon}_i$ are contractible.

The contracts the firm can offer are subject to the following standard restrictions. We assume that firm owners and workers are protected by limited liability and that contracts are monotone in cash flows, i.e., $0 \leq nw(n) \leq x$ and $0 \leq n\Delta w(n) \leq \Delta x$ for each n. The second (monotonicity) constraint ensures that both workers and firm owners are at least weakly better off in the high cash flow state. The motivation behind this assumption is that no party should have incentives to sabotage the firm in order to extract higher private payoffs (Innes 1990). In a similar vein, we restrict attention to contracts under which workers do not benefit when other workers leave, as that might hinder productive teamwork or even create incentives for mobbing or building a toxic working atmosphere.

Assumption 1 *For all shock realizations ε, a worker's expected on-the-job pay $W(\varepsilon, n) := w(n) + p(\varepsilon, n)\Delta w(n)$ is non-decreasing in n, i.e., a feasible contract must for all (ε, n) have a non-negative* retention sensitivity

$$\beta^n(\varepsilon, n) := W(\varepsilon, n+1) - W(\varepsilon, n) \geq 0. \tag{2}$$

We call a contract satisfying limited liability and monotonicity in both cash flow as well as the number of retained workers (see Assumption 1) a *feasible* contract. The firm can commit to any feasible contract at $t = 0$. All of the above is common knowledge.

a common practice in the U.S., where most workers are hired at will.

[12] While compensation that explicitly conditions on n might be rare in practice, contracts that implicitly condition on n are common. We discuss such contracts in detail within the context of the implementation of (optimal) contracts.

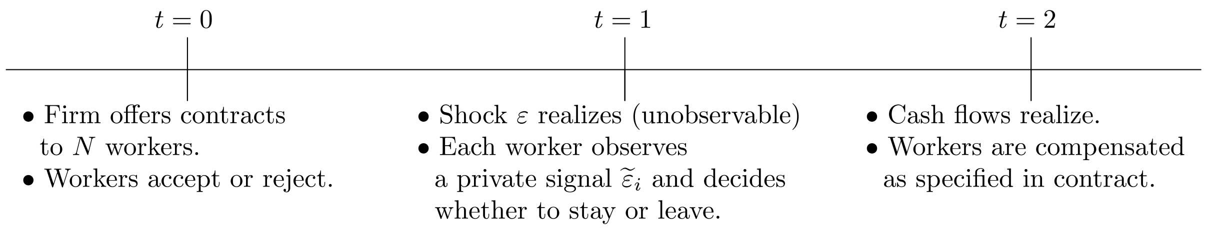


Figure 1: **Timeline.**

Our baseline model, summarized in Figure 1, makes several simplifying assumptions that we relax in subsequent sections. In particular, our main analysis assumes uniform contracting with commitment. In Section 5, we show that our results extend to more general compensation offers in which workers can choose from a menu of contracts based on their private observation of the interim signal, and we discuss the implications of renegotiation possibilities. Prior to that, in Section 4, we relax the assumptions that all workers are compensated with the same contract. We also consider the possibility that the firm can raise external financing to guarantee its workers' wages. While these extensions give rise to a number of additional predictions, they do not affect our main results.

3 Collective Turnover and Optimal Compensation

In what follows, we solve the model backward. In Section 3.1, we take the compensation contract as given and characterize the equilibria of the resulting coordination game at $t = 1$. Subsequently, in Section 3.2, we solve for the *optimal* profit maximizing compensation contract the firm offers at $t = 0$.

3.1 Workers' Coordination Problem

Take the compensation contract $C := \{w(n), \Delta w(n)\}_{n=1}^{N}$ as given and consider workers' decision of whether to stay with or leave the firm at $t = 1$. Each worker takes this decision based on a comparison of her expected on-the-job compensation and the available outside option. From Assumption 1, any given worker's expected payoff from staying is higher if more of her colleagues choose to stay as well. The fundamental reason is technological: the firm's success probability, $p(\varepsilon, n)$, increases in the number of workers the firm can retain, n, such that — for given payments in the low and high cash flow states — also expected on-the-job pay increases in n.[13] That is, workers play a coordination game with strategic

[13]This (positive) direct effect of an increase in n on expected compensation is a consequence of the assumption that $\Delta w \geq 0$. Assumption 1 effectively restricts a feasible contract's dependence on n such that

complementarities: an individual worker is more likely to stay if she believes that more of the other workers are staying as well. As is well known, such games may feature equilibria in which coordination fails. In particular, workers might choose to leave because they believe others are leaving, even though they would prefer to stay if everybody else stayed as well. We call a coordination failure of this form a "worker run."

3.1.1 Worker Runs and Compensation Remedies

To develop a basic understanding of the situations in which worker runs can arise, it is instructive to look at the benchmark in which the workers' signals perfectly reveal the (non-contractible) shock realization ε at $t = 1$ (i.e., when $\sigma = 0$). The following Proposition characterizes equilibria of the workers' coordination game in this case focusing on symmetric pure strategy equilibria.

Proposition 1 *Suppose that workers' signals perfectly reveal the shock ε at $t = 1$. Then, if*

$$W(\varepsilon, N) \geq \underline{w}(\varepsilon) > W(\varepsilon, 1), \tag{3}$$

there are two symmetric pure-strategy equilibria: In the first worker-run *equilibrium, all workers leave the firm and forgo their compensation in favor of their outside option. In the second equilibrium, all workers stay. Else, if condition (3) is violated, the workers have dominant strategies, and the equilibrium is unique: All workers stay if $W(\varepsilon, 1) \geq \underline{w}(\varepsilon)$ and all leave if $W(\varepsilon, N) < \underline{w}(\varepsilon)$.*

Proposition 1 identifies condition (3) as a necessary and sufficient condition for a worker run. A key implication is that the firm can avoid a worker run following a given shock realization, ε, by offering a contract under which workers' expected on-the-job compensation $W(\varepsilon, n) = w(n) + p(\varepsilon, n)\Delta w(n)$ is insensitive to the retention of other workers — i.e., $\boldsymbol{\beta}^{\mathbf{n}}(\varepsilon) = (\beta^n(\varepsilon, n))_{1 \leq n \leq N} = 0$ at this ε.[14] One way to achieve this is by offering a *fixed*-wage contract with $w(n) = w$ and $\Delta w(n) = 0$, under which $W(\varepsilon, n) = w$ for all n. Such a contract can ensure full retention whenever $w \geq \underline{w}(\varepsilon)$ if this is feasible.

The alternative is to offer a variable compensation contract with negative dependence on n, i.e., under which compensation in the low cash flow state, $w(n)$, or the high cash flow state, $w(n) + \Delta w(n)$, are decreasing in n (subject to maintaining Assumption 1). We refer to such compensation as "*dilutable*," as higher retention effectively dilutes the compensation

this direct effect is never dominated.

[14] We apply the standard convention that $\boldsymbol{\beta}^n(\varepsilon) = 0$ denotes $\beta^n(\varepsilon, n) = 0$ for all n. Accordingly, if $\boldsymbol{\beta}^{\mathbf{n}}(\varepsilon) = 0$ for all ε, which we denote by $\boldsymbol{\beta}^{\mathbf{n}} = 0$, worker runs can be ruled out for any realization of the productivity shock.

promised to workers. Dilutable pay reduces a contract's retention sensitivity by counteracting the positive effect of a higher retention level, n, on the workers' expected compensation. If appropriately designed, dilutable pay can completely offset this direct effect by reducing the retention sensitivity $\boldsymbol{\beta}^{\mathbf{n}}(\varepsilon)$ to zero. For example, by stipulating that $w(n) = 0$ and $\Delta w(n) = \frac{w(\varepsilon)}{p(\varepsilon,n)}$ for all n. With such a contract, changes in $p(\varepsilon, n)$ induced by n are perfectly offset by changes in $\Delta w(n)$ for a shock realization of ε. Intuitively, higher retention leads to a bigger pie, making it safe to dilute workers' share of the pie without threatening retention; on the other hand, worker departures cause the size of the pie to shrink, making it necessary to promise workers a bigger share of the pie. Standard implementations of such dilutable pay are through equity-based pay or bonus pools (see Section 3.2.3).

The currently considered benchmark setting with observable productivity shock ($\sigma = 0$) highlights worker runs as a coordination failure. In particular, whenever condition (3) holds such that the worker run equilibrium exists, there is a multiplicity of equilibria: workers stay with the firm if they believe that others are staying and leave if they believe that others are leaving. From an ex-ante ($t = 0$) perspective, a multiplicity arises whenever (3) holds for a subset of shock realizations as long as these have positive measure. For instance, dilutable compensation contracts that avoid runs for some shock realizations might not be able to do so for others. This multiplicity prevents a comparison of the entire set of feasible contracts in terms of their implied retention probabilities for all ε — a comparison needed for tackling the firm's optimal compensation design problem at $t = 0$. In order to resolve this problem, we turn to our main model specification in which workers observe the productivity shock with noise ($\sigma > 0$).

3.1.2 Quantifying the Probability of Retention

From now on, consider the main model specification in which each worker observes only an imperfect private signal $\widetilde{\varepsilon}_i$ about the productivity shock ε at $t = 1$, i.e., $\sigma > 0$. As we discuss next, this setting allows us to derive precise quantifiable predictions regarding a contract's retention features also in cases in which workers are not always (i.e., for all (ε, n)), better off staying or leaving. To do so, we distinguish two cases depending on whether a higher shock realization makes workers more or less likely to leave the firm under the given contract C. Which of these two cases arises is determined by the sensitivity of expected on-the-job compensation under C to the shock ε relative to that of the workers' outside option,

$$\beta^{\varepsilon}(\varepsilon, n) := \frac{\partial W(\varepsilon, n)/\partial \varepsilon}{\partial \underline{w}(\varepsilon)/\partial \varepsilon}. \tag{4}$$

We refer to (4) as the contract's *relative shock sensitivity.*

Consider, first, the case in which workers' on-the-job pay is more sensitive to productivity shocks than their outside option for all (ε, n) — i.e., when C has a shock sensitivity $\boldsymbol{\beta^{\varepsilon}} \geq 1$.[15] For example, this is always the case for idiosyncratic shocks that only affect the firm's productivity ($\partial p/\partial\varepsilon > 0$) but not the workers' outside options ($\partial \underline{w}/\partial\varepsilon = 0$). Whenever $\boldsymbol{\beta^{\varepsilon}} \geq 1$, staying becomes increasingly attractive to workers for higher shock realizations. Since higher signals $\widetilde{\varepsilon}_i$ are indicative of a higher realization of ε, we search for an equilibrium in which workers stay if and only if they receive a signal above some threshold. Workers leave for signals below the threshold, as that makes them pessimistic about the firm's probability of success and its ability to retain other workers instrumental for this success.

An equilibrium of this form exists if the following conditions are met (see, e.g., Morris and Shin, 2003). First, there are $\varepsilon', \varepsilon'' \in \mathbb{R}$ and $\delta > 0$ such that

$$W(\varepsilon'', 1) - \underline{w}(\varepsilon'') > \delta \text{ and } W(\varepsilon', N) - \underline{w}(\varepsilon') < -\delta, \tag{5}$$

with $\varepsilon'' > \varepsilon'$. This "dominance region" condition implies that workers are always better off staying with the firm if the shock is sufficiently high and leaving the firm if the shock is sufficiently low, regardless of the other workers' decisions to stay or leave. Outside of these dominance regions, a worker's optimal strategy again depends on her beliefs about how many of her colleagues are staying, which is increasing in the signal realization. The second condition we require is that there exists a unique cutoff ε^* defined as

$$\sum_{n=1}^{N} \frac{1}{N} \left(w\left(n\right) + p\left(\varepsilon^*, n\right) \Delta w\left(n\right) - \underline{w}\left(\varepsilon^*\right)\right) = 0. \tag{6}$$

Intuitively, ε^* is the signal that makes a worker indifferent between leaving or staying, given an (agnostic) uniform belief about the number of other players staying with the firm.[16] These conditions ensure that there is a unique cutoff equilibrium, in which a worker leaves if and only if she receives a signal lower than ε^*.[17] Any cutoff equilibrium of this form, thus, features a "worker run" for some realization of the workers' signals.

The second case we consider is when contracts have a shock sensitivity $\boldsymbol{\beta^{\varepsilon}} < 1$ — i.e., workers' on-the-job pay is less sensitive to shocks than workers' outside option for all (ε, n).

[15] Again, we apply the notational convention that $\boldsymbol{\beta^{\varepsilon}} \geq 1$ is equivalent to $\beta^{\varepsilon}(\varepsilon, n) \geq 1$ for all (ε, n).

[16] Given the assumptions on the signal structure with $\sigma \to 0$, posterior beliefs concerning the proportion of workers choosing to stay with the firm are almost uniform around the cut-off if everybody follows the same cut-off strategy.

[17] Equilibrium is unique in the sense that the described strategies are the only ones satisfying the iterated deletion of strictly dominated strategies (see, Proposition 2.2 in Morris and Shin 2003). While we established this result for the case in which the state is almost common knowledge, $\sigma \to 0$, this assumption is not necessary (see Morris and Shin, 2003, for a discussion).

This case can arise when the shock ε captures systematic risk that has a larger effect on the workers' outside options than on the firm, such as when $\partial p/\partial\varepsilon \to 0$, while $\partial \underline{w}/\partial\varepsilon > 0$. In this case, the workers' outside option is increasing faster in ε than their on-the-job pay, and higher signals make leaving the firm more attractive. We again assume that there exist $\varepsilon', \varepsilon'' \in \mathbb{R}$ and $\delta > 0$ such that condition (5) is satisfied — now with $\varepsilon'' < \varepsilon'$ — such that leaving is the dominant action for sufficiently high signals, while staying is dominant if the observed signal is sufficiently low. Then, given the existence of a unique solution ε^* to (6), there is a unique cutoff equilibrium in which a worker leaves the firm at $t = 1$ if she observes a signal $\widetilde{\varepsilon}_i > \varepsilon^*$ and stays with the firm if $\widetilde{\varepsilon}_i \leq \varepsilon^*$. The following Lemma summarizes these insights.

Lemma 1 *Suppose there is a unique ε^* solving condition (6). Then:*
(i) If $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} \geq 1$ and condition (5) is satisfied (with $\varepsilon'' > \varepsilon'$), worker i leaves the firm at $t = 1$ if she observes a signal $\widetilde{\varepsilon}_i < \varepsilon^*$ and stays with the firm if $\widetilde{\varepsilon}_i > \varepsilon^*$.
(ii) If $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} < 1$ and condition (5) is satisfied (with $\varepsilon'' < \varepsilon'$), worker i leaves the firm at $t = 1$ if she observes a signal $\widetilde{\varepsilon}_i > \varepsilon^*$ and stays with the firm if $\widetilde{\varepsilon}_i < \varepsilon^*$.

In the boundary case of $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} = 1$, a higher shock realization ε increases on-the-job pay and the outside option by the same amount. Lemma 1 applies again, with (3) holding either always – for all ε – or never. Together with the characterization in Lemma 1, this allows us to offer precise predictions regarding the probability of a run for essentially any "state monotonic" contract with $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} \geq 1$ or $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} \leq 1$. As it turns out, this will be sufficient to identify the main determinants of optimal compensation design.[18]

3.2 Optimal Compensation Design

We are now able to state the firm's compensation design problem. Since workers are ex-ante symmetric, we start by assuming that they are all hired with the same contract. We extend the analysis, allowing the firm to compensate identical workers differently in Section 4.1.

The firm's problem is to choose a feasible contract $C := \{w(n), \Delta w(n)\}_{n=1}^{N}$ to maximize expected cash flows net of compensation costs

$$\max_{C} \mathrm{E}_0 \left[x + p(\varepsilon, n) \Delta x - nW(\varepsilon, n)\right], \tag{7}$$

where E_0 denotes the expectation at $t = 0$ over the true shock realization ε and the signals $\widetilde{\varepsilon}_i$, $i = 1, ..., N$, that the workers observe at $t = 1$. For a given contract, these signals determine

[18] "State monotonicity" trivially holds for any contract in the firm's feasibility set if risk is idiosyncratic. In the case of systematic risk, this can be guaranteed by imposing economic restrictions on the firm's production technology.

the number of workers, n, that stay with the firm until $t = 2$. We assume throughout that the relevant set of feasible contracts is such that the coordination game at $t = 1$ has a unique equilibrium and provide sufficient conditions below. That is, worker i stays with the firm at $t = 1$ if and only if

$$\mathrm{E}_1^i \left[W(\varepsilon, n) - \underline{w}(\varepsilon) \,|\, \widetilde{\varepsilon}_i \right] \geq 0, \tag{8}$$

where E_1^i denotes the expectation at $t = 1$ given worker i's information set. Note that any feasible contract offer also has to satisfy limited liability and monotonicity: $0 \leq nw(n) \leq x$, $0 \leq n\Delta w(n) \leq \Delta x$, and $\beta^r(\varepsilon, n) \geq 0$ for each (ε, n). Furthermore, note that if the interim participation constraints in (8) is satisfied, so is the ex ante participation constraint.

3.2.1 First-Best Contracts and Main Determinants of Optimal Compensation Design

Given that higher retention increases the joint surplus of the firm and its workers, the upper bound for firm profits is achieved if (i) all N workers stay with probability one (efficiency) and (ii) workers can be pushed down to their reservation utility $\underline{w}(\varepsilon)$ for any signal realization (full rent extraction). This upper bound can be achieved if there exists a feasible contract under which the workers' expected on-the-job compensation perfectly matches her outside option,

$$w(n) + p(\varepsilon, n)\Delta w(n) = \underline{w}(\varepsilon), \ \forall (\varepsilon, n). \tag{9}$$

Taking the firm's perspective, we call such a contract a "first-best" contract. From (9) it holds that a first-best contract has a retention sensitivity of $\boldsymbol{\beta}^n = 0$ (since the outside option is constant in n) and a shock sensitivity of $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} = 1$ (since the contract needs to match the outside option's dependence on ε one-for-one).

In the following, we identify and characterize the two leading cases in which feasible first-best contracts exist. First-best compensation design in these cases illustrates the main determinants of the firm's optimal choice between fixed and dilutable pay as means of improving retention. The first case in which first-best is achievable is when risk is idiosyncratic and the firm has sufficient assets in place.

Proposition 2 *Suppose that shocks are idiosyncratic and the firm's assets in place satisfy $x \geq N\underline{w}$. Then, the firm can achieve its first-best payoff by offering a fixed wage of $w(n) = \underline{w}$ to all workers, while $\Delta w(n) = 0$ for all n.*

Notably, in the case of purely idiosyncratic risk, in which all contracts involving $\Delta w > 0$ have a (relative) shock sensitivity of $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} \to \infty$, a fixed wage of $\underline{w}$ is the only contract that can achieve first-best. However, the firm's resource constraint in the low cash flow state could make it impossible to promise such a fixed wage contract to all workers.

Instead, if risk is systematic, a pure fixed wage contract can no longer achieve first-best, as such a contract has a relative shock sensitivity of $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} < 1$. Thus, if a first-best contract exists, it must involve variable compensation (to achieve $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} = 1$). As a consequence, compensation needs to be dilutable such as to achieve a retention sensitivity of $\boldsymbol{\beta}^{n} = 0$. Since the shock ε is not contractible, satisfying (9) for all ε requires $\frac{\partial}{\partial \varepsilon} p(\varepsilon, n)$ to be proportional to $\frac{\partial}{\partial \varepsilon} \underline{w}(\varepsilon)$ whenever risk is systematic. As an illustration, we consider a specification in which $p(\varepsilon, n)$ and $\underline{w}(\varepsilon)$ are both linear in ε:

$$p(\varepsilon, n) = p_a(n) + \varepsilon p_b(n) \text{ and } \underline{w}(\varepsilon) = \underline{w}_a + \varepsilon \underline{w}_b, \tag{10}$$

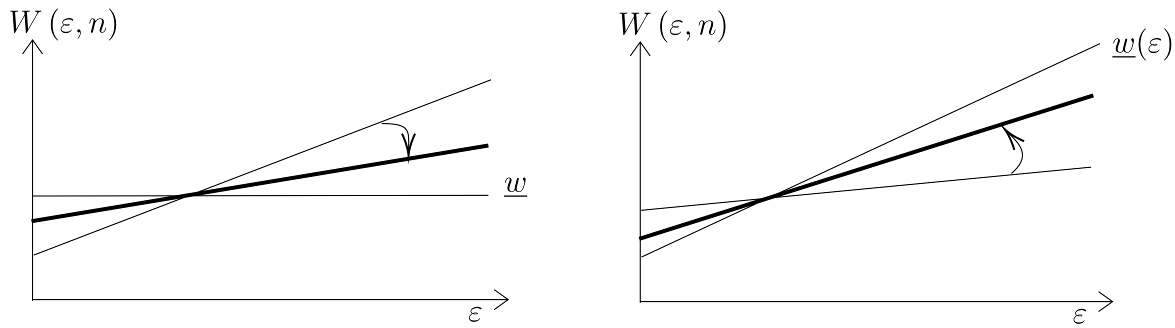
where $\varepsilon \in [0, 1]$, $p_a(1), p_b(1), \underline{w}_a, \underline{w}_b > 0$, and $p_a(n)$ as well as $p_b(n)$ increases in n with $p_a(N) + p_b(N) < 1$.

Proposition 3 *Suppose that shocks are systematic and* $p(\varepsilon, n)$ *and* $\underline{w}(\varepsilon)$ *are given as in (10). Then, the firm can achieve its first-best payoff by offering* $\{w(n), \Delta w(n)\}_{n=1}^{N} = \left\{\underline{w}_a - p_a(n) \frac{\underline{w}_b}{p_b(n)}, \frac{\underline{w}_b}{p_b(n)}\right\}_{n=1}^{N}$*, which is feasible if* $\underline{w}_a - p_a(n) \frac{\underline{w}_b}{p_b(n)} \in \left[0, \frac{x}{n}\right]$ *and* $\frac{\underline{w}_b}{p_b(n)} \leq \frac{\Delta x}{n}$ *for all* n.

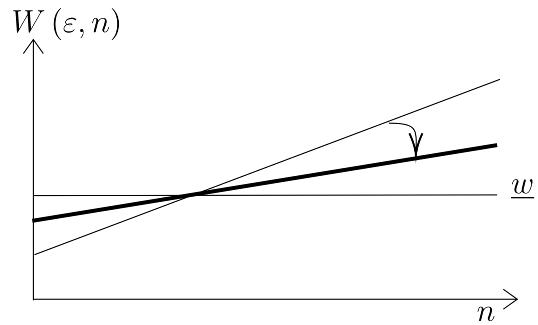
The key insight from Proposition 3 is that the first-best contract in the case of common (systematic) shocks relies on both fixed and dilutable variable pay. In particular, pay in the high cash flow state, $w(n) + \Delta w(n)$, is strictly decreasing in n.[19] The fixed non-dilutable part of compensation that workers receive in any case is given by $\min_n w(n)$. Furthermore, the firm's relative use of fixed and dilutable pay $w(n)/\Delta w(n)$ is increasing in the firm's shock sensitivity, $p_b(n)$, relative to that of workers' outside options, $\underline{w}_b$. Intuitively, if the firm is more sensitive to shocks than workers' outside options, matching the sensitivity of these outside options requires offering compensation with lower exposure to such shocks. That is, $\Delta w(n)$ must be low, and workers will be paid more in the low cash flow state ($w(n)$ must be high). In the limit, as we approach the case of an idiosyncratic shock, $w_b = 0$, the firm only offers fixed compensation (if it is feasible).

The above two cases in which first-best contracts are feasible illustrate the key economic intuition behind optimal compensation design given the possibility of worker runs: The choice between fixed and dilutable pay as means of improving retention is driven by how closely these instruments allow the firm to match the worker's outside option in order to reduce worker

[19] Though the effect of n on $w(n)$ is ambiguous, the negative dependence of $\Delta w(n)$ on n is always strong enough to keep expected compensation $w(n) + p(\varepsilon, n)\Delta w(n)$ constant in n.



Panel A: If $\boldsymbol{\beta}^{\boldsymbol{\epsilon}} > 1$, reduce sensitivity to ε. **Panel B**: If $\boldsymbol{\beta}^{\boldsymbol{\epsilon}} > 1$, increase sensitivity to ε.



Panel C: If $\boldsymbol{\beta}^{\mathbf{n}} > 0$, reduce sensitivity to n.

Figure 2: **Objectives of optimal compensation design: heuristic illustration.**

rents. The two first-best cases also illustrate why achieving first-best might be infeasible: (i) the non-contractibility of ε could make it impossible to perfectly match workers' pay with their outside options for every ε (e.g., when $\frac{\partial}{\partial \varepsilon} p$ and $\frac{\partial}{\partial \varepsilon} \underline{w}$ are not proportional under common shocks); or (ii) the firm's resource constraint may bind (e.g., when shocks are idiosyncratic and the firm has little liquid assets in place).

Whenever offering a first-best contract is not feasible, the firm's problem of maximizing its expected payoff (given by (7)) boils down to a trade-off between achieving higher efficiency by improving retention and minimizing the workers' rent, $W(\varepsilon, n) - \underline{w}(\varepsilon)$. This problem can be tackled in two steps, analogous to Grossman and Hart's (1983) approach to solving the principal-agent model. In particular, in the first step, one solves the firm's compensation design problem of minimizing compensation costs (worker rents) for any given level of retention, as characterized by the cutoff ε^* (see Lemma 1). As heuristically illustrated in Panels A and C of Figure 2, for a given ε^*, the intuitive objective is to match the retained workers' on-the-job pay more closely to their outside options for different levels of retention, n, and productivity shocks, ε. This compensation design problem will be the focus of the subsequent analysis. Given a characterization of optimal compensation contracts for given ε^*, the firm's profit-maximizing choice of ε^* is straightforward to determine.

3.2.2 Rent Extraction in Second-Best Contracts

The main qualitative insight that the firm's relative shock sensitivity plays a central role in shaping optimal compensation design remains valid also when first-best is not achievable. Firms more sensitive to shocks than workers' outside options optimally offer flat contracts, i.e., compensation with a high fixed and a low performance-dependent component ($w(n)/\Delta w(n)$ high). By contrast, firms with a low relative shock sensitivity do the opposite. They offer steep contracts with a low fixed and high performance-dependent component ($w(n)/\Delta w(n)$ low). This mimics the finding in Proposition 3 that in first-best contracts $w(n)/\Delta w(n)$ increases in $p_b(n)/w_b$. In order to illustrate this in the simplest possible way, we will focus our analysis on the cases in which all contracts in the firm's feasibility set, implementing an (interior) ε^* either satisfy $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} \geq 1$ or $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} < 1$.[20]

First, consider the case in which feasible contracts have $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} \geq 1$ (high relative shock sensitivity). A sufficient condition for this case to arise is that shocks are purely firm-specific (idiosyncratic), i.e., affect firm cash flows, $\frac{\partial}{\partial \varepsilon} p(\varepsilon, n) > 0$, $\forall n$, but not the workers' outside option, $\underline{w}(\varepsilon) = \underline{w}$. In this case, the first-best fixed-wage-only contract with $w(n) = \underline{w}$ and $\Delta w(n) = 0$ as characterized in Proposition 2 is infeasible whenever $x < N\underline{w}$. For ease of exposition, but without loss of generality for our main result, we impose the stricter resource constraint that $x < \underline{w}$. Since resources in the low cash flow state are insufficient to match workers' outside options, contracts under which the firm retains workers with non-zero probability must have a variable component $\Delta w(n) > 0$ for at least some n. Accordingly, all contracts in the firm's feasibility set that implement a given interior ε^* have $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}}(n) > 1$ for at least some n. Under such contracts, workers leave following signals smaller than ε^* and stay following signals larger than ε^* in which case they earn a rent (Lemma 1).

To reduce the workers' rent, the optimal second-best compensation design tries to get as close as possible to the first-best solution of offering a fixed-wage-only contract (Proposition 2). This is achieved by making compensation as flat as possible. That is, the firm can reduce the rent workers earn for $\varepsilon > \varepsilon^*$ by shifting as much compensation from the upside $\Delta w(n)$ to the fixed component $w(n)$ as is feasible. Doing so reduces the sensitivity of workers' expected compensation to the firm's probability of success, $p(n, \varepsilon)$, lowering the sensitivities to both the shock ε, $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}}(n)$, and the retention of other workers n, $\boldsymbol{\beta}^{\mathbf{n}}(n)$. As a result, the workers' on-the-job pay matches more closely their constant outside options.

Formally, for given ε^* — which fixes expected compensation at the cut-off ε^* according

[20] The case of a firm, which has contracts with both $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} > 1$ and $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} < 1$ in its feasibility set, is a straightforward extension. Contracts in which $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}}(\varepsilon) > 1$ for some ε and < 1 for others violate the state monotonicity assumptions necessary for the characterization of equilibrium in the induced coordination game workers play at $t = 1$.

to (6) — the firm optimally sets the ratio $w(n)/\Delta w(n)$ to its upper feasibility bound, which is defined by the firm's limited resources in the low cash flow state, $w(n) \leq \frac{x}{n}$, as well as the monotonicity constraints $\Delta w(n) \geq 0$ and $\boldsymbol{\beta}^{\mathbf{n}} \geq 0$ (see Assumption 1). Which of these upper feasibility constraints binds for which n and, thus, pins down the solution to the firm's linear programming problem depends on the precise shape of the set of feasible contracts, i.e., on parameters and functional form assumptions. Given such assumptions, the program can be solved numerically. The robust economic insights in high shock sensitivity firms ($\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} \geq 1$) are that: (i) compensation is as flat as possible, with $w(n)/\Delta w(n)$ determined by its upper feasibility bound; and (ii) compensation is dilutable, i.e., $w(n)$ or $w(n) + \Delta w(n)$ are decreasing in n. As in first-best, dilutable pay improves retention by paying remaining workers more when others are leaving.

Consider, next, the case in which feasible contracts have $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} < 1$. A sufficient condition for this case of a "low shock sensitivity firms" to arise is that the firm's expected cash flows are less sensitive to the systematic shock than workers' outside options

$$\frac{\partial}{\partial \varepsilon} \left(x + p(\varepsilon, n) \Delta x - n\underline{w}(\varepsilon) \right) < 0, \ \forall (\varepsilon, n). \tag{11}$$

Since $n\Delta w(\varepsilon) \leq \Delta x$, condition (11) implies that the rent, $W(\varepsilon, n) - \underline{w}(\varepsilon)$, a worker can earn in such a firm decreases in ε for all feasible contracts.

In this case, the problem faced by the firm reverses: Workers stay and earn rent in expectation following low shock realizations, while they are better off leaving following high shock realizations (Panel B in Figure 2). Thus, the main challenge for a firm with low shock sensitivity is to make on-the-job compensation attractive for workers observing high signals $\widetilde{\varepsilon}_i$ for which the outside option is relatively valuable, without overpaying workers (relative to their outside options) when their signals are low.

Compensation that resolves this problem needs to be sensitive to systematic shocks so as to track more closely workers' outside options, $\underline{w}(\varepsilon)$ (see Panel B of Figure 2 for an illustration). Offering workers a higher upside $\Delta w(n)$, while lowering their fixed payment, $w(n)$, achieves this goal and, thus, allows the firm to reduce workers' rent for any given level of retention. That is, contrary to the case in which $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} \geq 1$, the firm optimally offers the steepest possible contract, i.e., sets $w(n)/\Delta w(n)$ as low as possible such as to minimize the rent workers earn for $\varepsilon < \varepsilon^*$.

Formally, in low shock sensitivity firms ($\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} < 1$), $w(n)/\Delta w(n)$ is determined by its *lower* feasibility bound, which for given ε^* — fixing expected compensation at the cut-off ε^* according to (6) — is defined by limited liability $w(n) \geq 0$, and monotonicity constraints $\Delta w(n) \leq \frac{\Delta x}{n}$, and $\boldsymbol{\beta}^{\mathbf{n}} \geq 0$. This is again analogous to the first-best contract from Proposition 3, in which $w(n)/\Delta w(n)$ decreases towards zero as the relative shock sensitivity $\frac{p_b(n)}{w_b}$ decreases. The following Proposition summarizes these insights:[21]

Proposition 4 *Consider the firm's optimal compensation design problem of implementing a given level of retention ε^*, as defined by (6), at lowest compensation costs and suppose that $x < \underline{w}$.*
(i) If all contracts in the firm's feasibility set have $\boldsymbol{\beta^\varepsilon} \geq 1$, strictly so for at least one n, then the firm optimally offers the flattest feasible contract: $w(n)/\Delta w(n)$ is set to the maximal value allowed for by the constraints $w(n) \leq \frac{x}{n}$, $\Delta w(n) \geq 0$, and $\boldsymbol{\beta^n} \geq 0$ for the required retention level defined by (6). That is, under the optimal contract, $w(n)/\Delta w(n)$ cannot be increased for any n without decreasing it for another.
(ii) If all contracts in the firm's feasibility set have $\boldsymbol{\beta^\varepsilon} < 1$, then the firm optimally offers the steepest feasible contract: $w(n)/\Delta w(n)$ is set to the minimum value allowed for by the constraints $w(n) \geq 0$, $\Delta w(n) \leq \frac{\Delta x}{n}$, and $\boldsymbol{\beta^n} \geq 0$ for the required retention level defined by (6). That is, under the optimal contract, $w(n)/\Delta w(n)$ cannot be decreased for any n without increasing it for another.
Any optimal contract conditions on the number of retained workers and features dilutable pay.

Propositions 3 and 4 offer the same implication regarding whether firms optimally dull the strategic complementarities in workers' decision to leave ($\boldsymbol{\beta^n}$) via fixed (non-dilutable) or variable (dilutable) pay. While low-shock-sensitivity firms ($\boldsymbol{\beta^\varepsilon} < 1$) minimize the use of fixed (non-dilutable) pay, high-shock-sensitivity firms ($\boldsymbol{\beta^\varepsilon} \geq 1$) prefer the opposite. One implication is that firms with a lower sensitivity to systematic shocks (than workers' outside options) will offer more variable and less fixed pay. We should note that in a related single-worker setting, Oyer (2004) obtains that firms with higher exposure to systematic shocks should use less fixed and more equity compensation. Qualitatively, this prediction is opposite to ours. As discussed in detail in the Introduction, this has to do with the fact that Oyer assumes that all contracts must be less sensitive to shocks than workers' outside options in order to insure risk-averse workers. In our setting, this is suboptimal for the purpose of minimizing workers' rent extraction. Furthermore, it may not be technologically feasible or violate the firms' resource constraints.

[21] Throughout we assume that condition (5) of Lemma 1 is satisfied for feasible contracts implementing ε^*. A sufficient condition for the case of $\boldsymbol{\beta^\varepsilon} \geq 1$ is that $\inf_\varepsilon \Omega = 0$ and $\sup_\varepsilon \Omega = \Delta x - \underline{w}$. Similarly, for the case of $\boldsymbol{\beta^\varepsilon} < 1$, a sufficient condition is that $\inf_\varepsilon \underline{w}(\varepsilon) \leq 0$ while $\sup_\varepsilon \Omega = 0$.

3.2.3 Implementation of "Dilutable" Compensation

The preceding analysis highlights the importance of dilutable compensation, i.e., that compensation $w(n)$ in the low cash flow state or $w(n) + \Delta w(n)$ in the high cash flow state is decreasing in n. We acknowledge that compensation contracts that explicitly condition on n are rare in practice. However, there are various commonly used compensation components that implicitly contain the dilution features we show to be optimal in Propositions 3 and 4.

The prime example is equity-based compensation. In particular, many firms pay out part of their deferred compensation in terms of equity or stock options that vest only if the worker stays for a prespecified time, which in our model is till the end of $t = 2$. Since no (new) equity is issued to workers who leave the firm at $t = 1$, such an arrangement mechanically implies that remaining workers incur less dilution of their equity positions if other workers leave.[22] Compensation pools follow a similar principle: firms set aside a bonus pool as a given percentage of the firm's net revenues to be shared among employees at the end of a prespecified period. The more employees stay till the end, the lower the share each worker receives out of the pool. As another example, in partnerships, such as law or private equity firms, the accumulated bonus or carried interest is shared similarly. Overall, it is again important to note that dilution does not make workers worse off ex-ante since the higher level of retention makes it more likely that the firm is successful.

One potential issue with using equity-based pay to minimize the sensitivity, $\boldsymbol{\beta}^{\mathbf{n}}$, of workers' compensation to the retention of other workers is that equity-based pay is very sensitive to cash flow shocks, which — in the case of firm-specific shocks — creates a tension with the second objective of minimizing the sensitivity to such shocks, $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}}$. A way to mitigate this conflict is to offer workers equity ownership that can be diluted by non-employee owners in high cash flow states in which retention is not at risk — a common practice in start-ups.[23] In a nutshell, optimally exploiting the often-neglected dilution features of equity-based pay can make such compensation insensitive to the retention of other workers and cash flow shocks. This property underlines the appeal of using equity-based pay — instead of other forms of deferred pay — for retention purposes.

[22] Abstracting from our assumption of identical workers, commitment to internal promotions is another example of dilutable pay. While a worker might fear that the departure of her superior hurts firm profitability, the chance of being promoted to a better-paying job can mitigate the resulting incentives to leave.

[23] Concretely, the firm may issue (underpriced) equity to non-employee 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.

4 Asymmetric Compensation and External Financing

4.1 Compensating Identical Workers Differently

Our assumption that workers are identical and contribute the same to the value creation within the firm made it natural to assume that they receive the same contracts, which could be motivated, for example, based on a notion of "fairness." However, as we show next, the firm might achieve better outcomes by offering identical workers different types of compensation contracts. Since workers in practice are rarely identical, the key implication is that even marginal differences among workers (which create a marginally higher preference to retain some workers) could lead to large differences in compensation design. Crucially, this does *not* necessarily mean that those workers that are retained with higher probability are paid more than others, but rather that their compensation is structured differently. Indeed, it may even be optimal to allow workers that are retained with a lower probability to extract more rent.

To illustrate under which conditions and how asymmetric contracts can make the firm better off, it is instructive to look at two simple examples with purely idiosyncratic risk and a total of $N = 2$ workers. For both examples, assume that the firm's liquid assets in place are $x = \underline{w}$, such that the firm cannot achieve its first-best payoff when offering the same contract to both workers. Furthermore, the implementation of any given level of retention $\varepsilon^* < \bar{\varepsilon}$ with symmetric contracts requires that workers participate in the upside, $\Delta w(n) > 0$ for at least some n. Hence, the workers earn a strictly positive rent in expectation (see Proposition 4). In what follows, we illustrate two cases in which asymmetric contracts $\{w_i(n), \Delta w_i(n)\}$ for $i, n = 1, 2$ can improve on this outcome.

Example 1 In the first example, the main factor responsible for value creation is hiring and retaining one worker. That is, the surplus $\Omega(\varepsilon, 1)$ as defined in (1) is large, while hiring a second worker only has an incremental impact, i.e., $\Omega(\varepsilon, 2) - \Omega(\varepsilon, 1)$ is relatively small. Concretely, assume that the difference in success probabilities $p(\varepsilon, 2) - p(\varepsilon, 1) =: \Delta p$ is constant for all ε, and it holds that $\Delta p \Delta x = \underline{w} + \delta$, where δ is positive but small. By offering worker i a compensation contract $\{w_i(n), \Delta w_i(n)\} = \{\underline{w}, 0\}$ independently of n, this worker can be retained with probability one without paying her any rent. The contract offered to the other worker, $-i$, could then take the form $\{w_{-i}(2), \Delta w_{-i}(2)\} = \left\{0, \frac{\underline{w}}{p\left(\varepsilon^*_{-i}, 2\right)}\right\}$, where ε^*_{-i} is implicitly defined by $W\left(\varepsilon^*_{-i}, 2\right) = \underline{w}$.[24] That is, whenever worker $-i$ is retained with

[24] It is irrelevant how $\{w_{-i}(1), \Delta w_{-i}(1)\}$ is specified for worker $-i$, since worker i stays with probability one.

positive probability, $\varepsilon_{-i}^* < \overline{\varepsilon}$, she earns a rent. The firm will then optimally choose ε_{-i}^* such that worker $-i$'s expected rent is at most equal to her value-added, δ. If δ is small, these asymmetric contracts allow the firm to capture almost the entire surplus.

Example 2 In this example, the key aspect is that the firm creates a particularly high surplus, i.e., has a particularly high probability of success if it retains both workers. Concretely, assume that the probability of success is equal to one if both workers can be retained, $p(\varepsilon, 2) = 1$, and $\varepsilon > \underline{\varepsilon}$. Else, the probability of success is strictly less than one with $p(\underline{\varepsilon}, 1) = p(\underline{\varepsilon}, 2) = 0$, $\frac{\partial}{\partial \varepsilon} p(\varepsilon, 1) > 0$, and $p(\overline{\varepsilon}, 1) < 1$. In this setting, the firm can avoid runs by compensating the first worker with $\{w_i(n), \Delta w_i(n)\} = \{\underline{w}, 0\}$ and the second worker with $\{w_{-i}(n), \Delta w_{-i}(n)\} = \{0, \underline{w}\}$. With such contracts, it is a dominant strategy for worker i to stay, leading to a unique equilibrium in which both workers stay almost surely and extract no rent.[25] Again, the asymmetric contract strictly dominates any feasible symmetric contract.

The benefit of asymmetric compensation when firms are subject to resource constraints (and, thus, cannot offer safe compensation to everyone) stems from the following intuitive insights: By offering some workers safer compensation than under the second-best symmetric contract, the firm can retain this group of workers with a higher probability without having to grant them the same increase in rent as under the corresponding symmetric contract. This can be optimal, for example, if most value is created when a certain number of workers is retained (Example 1), or when guaranteeing the retention of some workers reduces the strategic uncertainty for others, making them more likely to stay (Example 2).

While prior work has shown that asymmetric compensation levels can sometimes help in mitigating coordination problems (Winter, 2004), our focus is on the role of asymmetric compensation structure. That is, the novel question we are after is how the different contracts offered to different groups of identical workers (that differ only in the probability of retention the firm wants to induce for each group) are optimally structured. In line with Examples 1 and 2, we will focus on the case in which $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} \geq 1$. This case is particularly interesting from a corporate finance perspective, as the firm then needs to decide to which workers to allocate its limited resources in the low cash flow state.

4.1.1 Optimal Asymmetric Compensation

In what follows, we refer to the group of workers the firm wants to retain with a higher probability as group 1 and the remaining workers as group 2. It is optimal to compensate group 1 with a contract with a lower sensitivity to shocks than the one offered to group 2.

[25] Worker $-i$ is better off leaving if and only if $\varepsilon = \underline{\varepsilon}$, which is a zero probability event.

That is, similar to the examples above, group 1 will receive more fixed compensation, $w(n)$, and less variable pay, $\Delta w\,(n)$.

The trade-off behind this result is that offering group 1 compensation with higher fixed pay mechanically implies that the firm cannot promise as much fixed pay to group 2 due to binding resource constraints. As a result, reducing the rent that group 1 can extract comes at the expense that (ceteris paribus) group 2 will be able to extract *more* rent: Looking at Panel A of Figure 2, group 1's expected payoff becomes flatter while that of group 2 becomes steeper. The reason the firm optimally offers group 1 the flatter contract is that this group stays with a higher probability than group 2. Thus, reducing group 1's rent is the more cost-efficient choice.

To formalize these insights, we use the following notation: Denote the set of workers in group 1 (which are retained with higher probability) by $N_1 \subset \{1, ...N\}$ and the number of workers retained from this group by n_1. Similarly, the number of workers retained from $N_2 = \{1, ...N\}/N_1$ (group 2) is given by n_2. We consider symmetric contracts within each group, i.e., for all $\mathbf{n} = \begin{pmatrix} n_1 & n_2 \end{pmatrix} \in N_1 \times N_2$ workers are compensated according to contract $\{w_1(\mathbf{n}), \Delta w_1(\mathbf{n})\}$ if they are from group N_1 and according to $\{w_2(\mathbf{n}), \Delta w_2(\mathbf{n})\}$ if they are from group N_2.[26] Examples 1 and 2 illustrate that feasible asymmetric contracts exist. The following proposition characterizes these contracts by showing that it is optimal to shift compensation in the low cash flow state from group N_2 to group N_1 for any given $\mathbf{n}$ until either resource or monotonicity constraints bind.

Proposition 5 *Suppose feasible contracts have a shock sensitivity greater than one (i.e., $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} \geq 1$) and that the firm wants to increase the retention probability of a subset of workers $N_1 \subset \{1, ...N\}$. Then, if an equilibrium in cutoff strategies exists in which the cutoff ε_1^* for workers $i \in N_1$ is less than the cutoff ε_2^* for workers $j \in N_2$, the ratio $w_1(\mathbf{n})/\Delta w_1(\mathbf{n})$ for workers in group N_1 is set to the maximal value allowed for by the resource and monotonicity constraints, $n_1 w_1(\mathbf{n}) + n_2 w_2(\mathbf{n}) \leq x$, $\Delta w_1(\mathbf{n}) \geq 0$, $\Delta w_2(\mathbf{n}) \geq 0$, and $\boldsymbol{\beta}^{\mathbf{n}} \geq 0$, as well as the required retention levels ε_1^* and ε_2^*. That is, under the optimal contracts, $w_1(\mathbf{n})/\Delta w_1(\mathbf{n})$ cannot be increased for any $\mathbf{n}$ without decreasing it for another.*

4.1.2 Sequential Departures and Key-Man Risk

Our baseline model assumes that workers simultaneously decide whether to leave the firm. While this is a theoretical simplification, the relevant assumption is that when taking her

[26] That is, we assume that contractual payments at $t = 2$ can condition on the realized cash flow state and the number of workers that have been retained from each group separately. Since we assume that all compensation is deferred, we immediately have that $w_1(\mathbf{n}) = 0$ for all $\mathbf{n} = \begin{pmatrix} 0 & n_2 \end{pmatrix}$ and similarly $w_2(\mathbf{n}) = 0$ for all $\mathbf{n} = \begin{pmatrix} n_1 & 0 \end{pmatrix}$.

decision whether to stay with or leave the firm, each worker is unaware of other workers' decisions.[27] We now discuss an alternative setting in which workers take decisions sequentially, and decisions are directly observable.

In particular, assume that it is commonly known that one employee (the "initial" worker) receives a signal about the firm's prospects and the value of an outside offer before all other workers, e.g., because she is in a leadership position, or has skills attractive to a competitor. If it is optimal for the initial worker to take the decision to stay or leave immediately upon observing the private signal, e.g., because waiting erodes her outside option, then this generates two informational externalities: First, the fact that the initial worker's decision is observed, reduces strategic uncertainty for the remaining workers when taking their decisions simultaneously at $t = 1$, since they can be sure that at least one worker stays (leaves). Second, the remaining workers update their beliefs about the firm's prospects based not only on their signal but also on the initial worker's decision to stay or leave (as in an information cascade).

Offering the initial worker a tailored contract under which she is more likely to stay becomes now a natural choice. Even though the initial worker is equally productive to all others, the risk of her departure effectively becomes a "key-man risk": If that worker leaves, the firm's productivity drops with certainty, which increases the likelihood that all other workers leave.[28] Hence, such "key-man risk" is another motivation for the use of asymmetric compensation contracts as characterized in Proposition 5.

4.2 Effect of External Financing on Retention

To study the effect of external financing, we extend our model by stipulating that the firm needs to invest a capital outlay of K to start the project. We assume throughout that K is small enough such that the firm wants to invest. For ease of presentation, we further assume that K has to be raised externally from a perfectly competitive financial market.[29] The firm can further use its access to external financing to fund part of its wage obligations. The key novel implication from our subsequent analysis of the joint compensation and security design problem is that the optimal financing strategy depends on the firm's relative shock

[27] This is particularly realistic if the outside option is only available for a short period of time such that workers have to decide quickly, leaving little time for communication. Furthermore, firms often ask workers not to communicate that they handed in their resignation such that their departure only becomes observable to other workers at the end of the notice period.

[28] This direct effect of reducing strategic uncertainty, pushing towards contracts that retain the initial worker with higher probability, is likely to dominate the ambiguous information cascade effect.

[29] That is, we assume that the firm does not use its assets in place x to co-finance the investment. Since allowing for such co-financing does not affect our results but complicates the presentation, we abstract from it. Note that one may also view x as the cash flow in case of failure or the firm's liquidation value, in which case it is not available for investment purposes at $t = 0$.

sensitivity. While equity financing is preferable if $\boldsymbol{\beta}^{\varepsilon} \geq 1$, raising debt financing can help improve retention at a lower cost in firms with relatively low sensitivity to the productivity shock ($\boldsymbol{\beta}^{\varepsilon} < 1$). In order to highlight the effect of external financing in the simplest possible setting, we focus on symmetric compensation contracts.

Concretely, assume that the firm makes a take-it-or-leave-it offer to financiers, together with the offer it makes to workers. An external financing contract $S := \{s, \Delta s, I\}$ stipulates transfers to the financiers at $t = 2$ of s in the low cash flow state and of $s + \Delta s$ in the high cash flow state in exchange for an initial investment $I \geq K$. While we do not allow $\{s, \Delta s\}$ to depend on n, this is without loss for our analysis in what follows. The financing contract is commonly observable. All parties are risk-neutral, do not discount future payoffs and are protected by limited liability. As is standard, we restrict attention to monotone contracts, which ensures that no party has incentives to sabotage the project in the high cash flow state (Innes, 1990; Nachman and Noe, 1994). Formally, it should hold that $s, \Delta s \geq 0$ as well as $0 \leq nw(n) + s \leq x + I - K$ and $0 \leq n\Delta w(n) + \Delta s \leq \Delta x$. As in the baseline model, compensation contracts need to be monotone in n, i.e., $\boldsymbol{\beta}^{\mathbf{n}} \geq 0$ (Assumption 1). We refer to contracts satisfying these conditions as feasible.

The firm chooses compensation and security design to maximize its expected payoff

$$\max_{C,S} \mathrm{E}_0 \left[x + I - K - s - nw(n) + p(\varepsilon, n)(\Delta x - \Delta s - n\Delta w(n))\right], \tag{12}$$

subject to the financiers' ex-ante break-even constraint

$$\mathrm{E}_0 \left[s + p(n, \varepsilon) \Delta s\right] = I \tag{13}$$

the workers' interim participation constraints

$$\mathrm{E}_1^i \left[w(n) + p(n, \varepsilon) \Delta w(n) - \underline{w}(\varepsilon) \,|\widetilde{\varepsilon}_i\right] \geq 0, \tag{14}$$

and the feasibility restrictions on $\{s, \Delta s\}$ and $\{w(n), \Delta w(n)\}$ stated above. Analogous to the baseline model, the firm's objective is to optimally resolve the trade-off between achieving higher retention (and, thus, efficiency) and reducing compensation costs. The main addition to the baseline model is that the financing contract affects the feasibility bounds on workers' compensation. Through this effect, financing structure affects the retention – rent extraction trade-off. To illustrate the intuition behind the optimal joint design of compensation and financing, we distinguish again between the cases of high and low shock sensitivity firms.

First, consider a firm exposed to purely idiosyncratic risk and assume that the firm raises just enough to fund the initial investment, i.e., $I = K$. If the first-best compensation contract is not feasible since $x < N\underline{w}$, compensation contracts are characterized by Proposition 4. That is, the firm optimally reduces the sensitivity of workers' compensation both to retention and shocks by increasing the pay to workers in the low cash flow state, $w(n)$, subject to monotonicity and (crucially for the subsequent argument) the resource constraints $w(n) \leq x/n$. This resource constraint can be relaxed by external financing in excess of K. Specifically, by raising $I = K + N\underline{w} - x$, the firm can achieve full retention without paying any rents to workers by specifying $w(n) = \underline{w}$ for all n. Thus, we obtain that external financing makes the first-best compensation contract from Proposition 2 feasible. To facilitate such relaxation of the firm's resource constraint, optimal security design should shift all payments to investors to the high cash flow states, i.e., $s = 0$ and $\Delta s = \frac{K+N\underline{w}-x}{\mathrm{E}_0[p(N,\varepsilon)]}$. Raising more than $K + N\underline{w} - x$ is not optimal, as it is neither needed for investment nor relaxes any further constraints.

While the first-best outcome facilitated by external financing is only attainable in the case of pure idiosyncratic risk ($\underline{w}(\varepsilon) = \underline{w}$), the general intuition is robust. High shock sensitivity firms ($\boldsymbol{\beta^{\varepsilon}} \geq 1$) can relax relevant resource constraints preventing them from offering higher fixed wages, $nw(n) \leq x$, by securing external financing for such wages at $t = 0$. Investors are then only repaid once wages are paid in full. Issuing equity is a natural implementation.[30]

The opposite insights apply if the firm is less sensitive to productivity shocks than workers' outside option ($\boldsymbol{\beta^{\varepsilon}} < 1$). In this case, matching workers' on-the-job pay more closely to their outside options requires shifting compensation to the high cash flow state (Proposition 4). To facilitate this solution, the firm needs to preserve more of the upside, Δx, for compensation purposes, by shifting the payments to investors as much as possible to the cash flow-independent component s. Specifically, increasing s allows the firm to satisfy the investor's participation constraint (13) with a lower Δs, relaxing the monotonicity constraint $n\Delta w(n) + \Delta s \leq \Delta x$. Hence, in this case, the optimal security design resembles (risky) debt, where investors are repaid out of firm cash flows net of wages. The firm does not raise more than K, as that would only tighten the latter constraints.[31]

Proposition 6 *A high shock sensitivity firm optimally raises* $I = K + N\underline{w} - x > K$ *by*

[30] An alternative interpretation of this solution is that the firm only raises $I = K$ from external investors in exchange for $s = 0$ and $\Delta s = \frac{K}{\mathrm{E}_1[p(N,\varepsilon)]}$ and additionally enters an "insurance" contract that pays the firm $N\underline{w} - x$ in the low cash flow state in return for a premium of $\frac{N\underline{w}-x}{\mathrm{E}_1[p(N,\varepsilon)]}$ in the high cash flow state. In practice, a credit line can play the role of such insurance as could be formalized in a dynamic extension of our model.

[31] This can be formalized as follows: From Proposition 4 the optimal compensation contract for the $\boldsymbol{\beta^{\varepsilon}} < 1$ case sets $w(n)$ as low as feasible. However, this requires that resources in the high cash flow state are sufficient to pay workers enough such as to satisfy their participation constraint for all $\varepsilon \leq \varepsilon^*$, i.e., $\mathrm{E}_1^i[x - s + p(n, \varepsilon^*)(\Delta x - \Delta s) - \underline{w}(\varepsilon)] \geq 0$. This interim resource constraint is relaxed when K and, thus, Δs are lower.

issuing equity. A low shock sensitivity firm raises $I = K$ by issuing debt.

5 Discussion and Extensions

Compensation Design and Cash Flow Dispersion. Our analysis so far highlighted the importance of the sensitivity of workers' on-the-job pay to productivity shocks *relative* to the respective shock sensitivity of their outside options ($\boldsymbol{\beta}^{\varepsilon}$). A natural question then is how optimal compensation design is affected by an increase in the *level* of risk, which we model in terms of a higher cash flow dispersion in the sense of a second order stochastic dominance shift. Concretely, suppose that by investing $I < x$ and, thus, keeping liquid assets of $y := x - I$ (corresponding to available cash in case of failure), the firm could generate a higher upside in case of success, $\Delta y > \Delta x$. Assume that the increase in cash flow dispersion is mean-preserving, i.e., both projects have the same ex ante expected payoff conditional on retaining all workers:

$$\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \left(x - y + p\left(\varepsilon, N\right)\left(\Delta x - \Delta y\right)\right) dG\left(\varepsilon\right) = 0.$$

A higher cash flow dispersion tightens the firm's resource constraints in the low cash flow state since the funds available for compensation are lower, $y < x$. As a result, firms with $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} \geq 1$ that optimally rely on fixed pay $w(n) = \frac{x}{n}$ to achieve a given level of retention are worse off. In particular, they are forced to reduce $w(n)$ and rely more on variable pay, which allows workers to extract more rent. This is particularly, apparent in the idiosyncratic risk case when first-best is attainable under low cash flow dispersion, as $x \geq N\underline{w}$, but no longer possible if dispersion is higher as $y < N\underline{w}$ (Propositions 2 and 4). The effect for $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} < 1$ firms is, in general, less detrimental. Since such firms optimally rely mainly on variable compensation $\Delta w(n)$, they are less affected by the tightening of the resource constraint in the low cash flow state (unless a binding monotonicity constraint ($\boldsymbol{\beta}^{\mathbf{n}} \geq 0$) forces also such firms to offer $w\left(n\right) > 0$). In turn, the greater reliance on variable pay $\Delta w\left(n\right)$ means that such firms also potentially benefit more from the relaxation of the monotonicity constraint in the high cash flow state from $\Delta w\left(n\right) \leq \frac{\Delta x}{n}$ to $\Delta w\left(n\right) \leq \frac{\Delta y}{n}$.

Fixed Wages and Debt Financing. A possible concern with offering fixed wages to reduce the risk of worker runs (as optimal for high shock sensitivity firms) is that fixed wages can be seen as a form of debt. Taking this view, fixed wages increase the firm's operating leverage and make it potentially more prone to default. In turn, the higher risk of default could make workers more likely to run. The reason this argument fails is that it neglects that fixed wages have a higher priority in default than equity and are, thus, more valuable to workers when the firm's cash flows are low.

Another potential concern is that fixed wages could create a sort of a debt overhang problem, where the firm cannot raise new funding needed to be successful because new investors fear that a large portion of cash flows will be redirected to paying workers' old wage obligations. However, while wages are similar to debt, firms typically do not have to seek consent from workers before raising new debt with higher priority than wages. Thus, the risk of a "wage overhang" is easily overcome by raising new more-senior debt.[32]

Menus of Contracts. Our baseline model assumes away the possibility of offering menus to workers at $t = 0$. A natural question is whether the firm could do better by offering a menu from which the workers can choose different contracts depending on their private signals at $t = 1$.[33] This is not the case for any cutoff equilibrium in which workers stay if and only if they observe a signal above (or alternatively below) a certain threshold as in Lemma 1.

To see this, suppose that the shock is idiosyncratic, i.e., $w(\varepsilon) = 0$, and so $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} \geq 1$. Consider any non-degenerate menu W associated with an equilibrium cutoff strategy ε^*. Let $\widetilde{\boldsymbol{\omega}} = \{\widetilde{w}(n), \Delta\widetilde{w}(n)\} \in W$ be the contract chosen by the worker when her signal is ε^*. Now drop all contracts other than $\widetilde{\boldsymbol{\omega}}$ from the menu. Since $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} \geq 1$, if a worker prefers $\widetilde{\boldsymbol{\omega}}$ to her outside option $\underline{w}$ for some signal ε^*, the same holds for all higher signals $\varepsilon > \varepsilon^*$. Thus, the set of signals, Ω_{W_θ}, for which the worker stays at the firm remains unchanged. But then, by revealed preference of the worker for contracts other than $\widetilde{\boldsymbol{\omega}}$, the firm must be better off dropping these contracts. In particular, if there existed a contract $\boldsymbol{\omega} \in W_\theta$ that the worker strictly prefers over $\widetilde{\boldsymbol{\omega}}$ after a certain signal observation — and that she would, hence, choose from the menu since her expected payoff under $\boldsymbol{\omega}$ is higher than under $\widetilde{\boldsymbol{\omega}}$— this would necessarily imply that the firm's residual payoff for that signal is lower. Since the same argument applies to all workers that stay, the firm is better off offering only contract $\widetilde{\boldsymbol{\omega}}$. The argument that no menu can make the firm better off if $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}} < 1$ is analogous.

Renegotiation. The baseline model further assumes that the firm can commit to the contracts it offers and does not renegotiate with workers at $t = 1$. A complete analysis of renegotiation with multiple privately informed agents is beyond the scope of this paper. However, two extreme cases illustrate the key role of the firm's information at $t = 1$ in

[32]In such an extension of our model, workers will naturally anticipate that the firm may issue debt with higher priority. Still, ex ante, the higher expected investment efficiency will benefit all parties.

[33]Complex menus that condition each worker's contract on the reports sent by all other workers are unrealistic in the considered application and hence not considered.

determining whether contracts are renegotiation-proof.

In the first case, the firm does not observe the workers' signals at $t = 1$. The contract offered at $t = 0$ is then renegotiation-proof. In particular, since the firm's information does not change from $t = 0$ to $t = 1$, the same argument showing that it is not optimal to commit to a self-selection menu at $t = 0$ explains why renegotiating the original contract at $t = 1$ cannot be optimal. As a second extreme case, assume there is only one worker ($N = 1$), and the firm and the worker observe the same signal $\widetilde{\varepsilon}$. The key difference to the previous case is that the worker and the firm share the same posterior beliefs. Thus, the firm will renegotiate to retain the worker if the common posterior belief indicates that a worker's expected on-the-job pay is less than her outside option. Thus, when information is symmetric, renegotiations can achieve retention. Overall, we expect reality to be somewhere in between, with firms refusing to renegotiate if their lack of information makes it hard to distinguish whether increasing workers' pay is necessary to improve retention or just increasing rents; and renegotiating if they have information indicating that increasing workers' pay is necessary for retention.

Optimal hiring and optimal retention. The main focus of our analysis is on the compensation design implications of a firm's problem of retaining multiple workers. In particular, our analysis of the second-best case, in which full retention and surplus extraction cannot be achieved simultaneously, studied contracts implementing a given level of retention, as implied by the cutoff ε^*, at the lowest compensation costs. While we did not explore the optimal choice of ε^*, the main drivers of this decision are standard: the firm trades off the efficiency increase of higher retention against the required incremental rent. Since retention in our model is directly linked to (workers' signals about) productivity shocks a key determinant of the optimal choice of ε^* is whether the number of retained workers and productivity shocks are substitutes or complements in the firm's production technology, $p(\varepsilon, n)$.

Related, to focus on the problem of efficient retention, we took the number of workers, N, that the firm hires at $t = 0$ as exogenously given and assumed that it is efficient to retain as many as possible. A potentially interesting extension is how many workers the firm should hire at $t = 0$, anticipating that some of these workers will leave or that the optimal size of the workforce might be smaller for some shock realizations at $t = 1$. Without any cost of hiring, training, or laying off employees, the optimal policy is to hire as many workers as possible at $t = 0$.[34] However, in the presence of such costs, there will be a trade-off. The firm can either hire fewer workers at higher wages for which each individual worker is more

[34]This policy increases the probability that a sufficiently high number of workers will observe a signal for which they will stay even if the firm chooses ε^* to be very high (if $\boldsymbol{\beta}^\varepsilon > 1$) or low ($\boldsymbol{\beta}^\varepsilon < 1$), respectively, leaving little rent to workers. If the number of employees that want to stay at $t = 1$ is subsequently higher than what the firm needs, it will then lay off some of them.

likely to stay or more workers at lower wages for which each individual worker is less likely to stay, with the firm possibly laying off workers at $t = 1$ if too many want to stay. Notably, regardless of the strategy chosen by the firm, for any given number of workers the firm wants to retain, our analysis of the most cost-efficient way of doing so continues to apply.

Endogenous outside options. Our analysis focuses on a partial equilibrium (single firm) setting in which the workers' outside option is exogenously given. This is a realistic assumption in fragmented markets or in the case of startups, whose compensation choices are unlikely to affect those of established firms competing for the same workers.[35] If, instead, the workers' outside options endogenously adjust to the contracts offered by the firm, our previous results suggest the following: If a firm believed that other firms at which its workers could find alternative employment offer compensation with low sensitivity to productivity shocks, it would also be optimal for the firm to offer compensation with lower sensitivity to such shocks. By contrast, if the firm believed that other firms offer compensation with a high sensitivity to these shocks, it would also be optimal to also choose compensation with high sensitivity (Proposition 4).[36]

6 Empirical Implications

The annual quit rate among high-skilled workers is currently at 29% (BLS, 2021) and costs U.S. firms more than a trillion dollars a year (Gallup, 2019; Work Institute, 2019). When workers quit, they typically do so in waves, which is a first-order problem for firm performance (Felps et al., 2009; Hausknecht and Trevor, 2011; Hancock et al., 2013; Heavey et al., 2013). While worker departures can, in some cases, be socially efficient and may lead to better matching between workers and firms, our analysis focuses on the opposite case, in which contagious turnover erodes a firm's prospects and erodes efficient worker-firm matches.

Inefficient turnover is particularly likely to be a problem for firms that rely heavily on well-functioning teams of high-skilled workers with complementary skills. Keeping these teams together is then of first-order importance. Particularly at risk are firms in which departing workers are hard to replace both from within the team — e.g., because the team is small and concentrated with skills tied to individual workers — or from the outside — e.g., because the job requires specialized or firm-specific knowledge and the labor market is tight.

[35] Alternatively, the firm's main competitors may be subject to binding regulation on the structure of pay, e.g., as implemented for large banks in the form of deferral/clawback clauses or caps on bonuses (see, e.g., Hoffmann et al. 2021).

[36] Van Wesep and Waters (2021) study a related coordination problem in the context of firms choosing the optimal time to pay out bonuses.

In these cases, replacing workers with equally efficient ones on short notice is hard due to the lack of skill standardization (see Rajan 2012). High-growth startups, for which conventional wisdom dictates that "an A team with a B idea is more important than a B team with an A idea" are a primary example since their human capital is usually less standardized than in mature public firms (Rajan 2012). Other examples in which retaining teams of high-skilled workers is of utmost importance include firms in consulting, advisory, investment banking, law firms, and PE partnerships.[37] Apart from these classifications based on industry and maturity, empirical measures of team importance could be constructed, for example, by scraping job postings for phrases related to the importance of teamwork. Proxies for how difficult it is to replace workers could be constructed based on data about how fast firms can fill job vacancies. Overall, our compensation design predictions should apply in particular to firms for which worker runs are a relevant (cost) factor.

Implication 1 *Firms whose human capital consists mainly of teams of hard-to-replace high skilled workers pay a higher share of total compensation via fixed compensation or dilutable pay — such as claims on an equity stock option or a bonus pool.*

The underlying reason for this prediction is that firms in which the worker run problem is relevant, such as startups, can improve worker retention for a given level of total compensation by using fixed wages or dilutable (equity) compensation rather than, e.g., fixed performance bonuses. The popularity of using deferred equity-based pay for retention (Aldatmaz et al., 2018; Jochem et al., 2018; Hochberg and Lindsey, 2010) as opposed to other types of deferred (fixed) compensation is seen as suboptimal by prior work, as it exposes workers to risk beyond their control (Murphy, 2003; Lazear, 2004). From this perspective, the predominant use of equity-based pay in startups is particularly puzzling (Hand, 2008). Implications 1 sheds light on this discussion by pointing to the often-neglected dilution feature of equity-based pay. This feature effectively promises remaining workers higher pay when others are leaving but allows firms to dilute workers when retention is not at risk. Such dilution *reduces* the sensitivity of workers' compensation to risk, which helps cash-constrained firms get closer to the dual objectives of improving retention and minimizing workers' rents.

This dilution property of equity-based pay also sheds light on the structure of partnerships, such as law and private equity firms. For example, when a partner in a PE firm leaves, the accumulated carried interest she forgoes is distributed among remaining partners. Sim-

[37] 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.

ilarly, in law firms, departing partners typically have to return or sell back their equity to the firm, effectively increasing the equity ownership of remaining partners.

Our analysis further shows that the optimal mix of fixed and dilutable pay in addressing the risk of contagious turnover differs depending on whether the firm is more or less sensitive to (systematic) risk than workers' next best outside option (Propositions 3 and 4).

Implication 2 *The ratio of fixed to dilutable (equity-based) pay increases in firms' sensitivity to systematic shocks compared to that of workers' outside options. This relationship is particularly strong in firms heavily dependent on teams of hard-to-replace workers.*

Tests of Implication 2 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. Typically, these can be found in closely related firms operating in the same industry or requiring a similar skill set. Identifying such similar firms is crucial to adequately capture our notion of "relative shock sensitivity ($\boldsymbol{\beta}^{\boldsymbol{\varepsilon}}$)." In particular, even if a firm is weakly exposed to systematic shocks, in the sense of a low market beta, it need not have a low shock sensitivity $\boldsymbol{\beta}^{\boldsymbol{\varepsilon}}$. The reason is two-fold: First, closely related firms could have an even lower market beta. Second, these firms might rely heavily on fixed pay components, which effectively shield their workers from equity risk. One possibility to identify closely related firms is by using Hoberg and Phillips (2016) network similarity scores. Equity-based pay below the executive level could be measured as in Bergman and Jenter (2007). Using these measures, we find in unreported regressions support for both Implications 1-2.

Another perspective on Implication 2 is to look at different groups of workers within an organization that differ in the sensitivity of their respective outside options to shocks. In particular, workers with more generic skills, such as mid-level managers, are likely to be less sensitive to industry-specific shocks than workers with firm and industry-specific skills, such as engineers or scientists. Taking this perspective, our model would predict that workers with more generic skills will be compensated with more equity-based and less fixed pay compared to workers with firm- or industry-specific skills.

Implication 2 discusses the impact of a firm's (relative) sensitivity to risk compared to workers' outside options. However, the level of risk faced by the firm also matters (Section 5):

Implication 3 *Higher cash flow dispersion increases the fraction of dilutable (equity-based) pay to fixed compensation in financially constrained firms with a high sensitivity to (systematic) shocks relative to workers' outside options. In contrast, there is little effect of such an* *increase in risk on compensation design when firms are unconstrained or have a low relative shock sensitivity.*

Our paper further provides clear implications on how the association between firm leverage and retention depends on firms' sensitivity to risk (see Proposition 6):

Implication 4 *(i) Firms with low sensitivity to (systematic) shocks compared to workers' outside options can improve retention for a given level of compensation costs by financing their investments with debt and preserving equity for compensation purposes. (ii) By contrast, firms with a high relative sensitivity to systematic or idiosyncratic risk will use equity financing or credit lines.*

Prior theory has predicted an unambiguously negative relation between higher leverage and turnover (Titman, 1984; Berk et al., 2010). However, empirical work has found mixed evidence for this prediction, with some studies finding supportive evidence (Chemmanur, Cheng, and Zhang, 2013; Agrawal and Matsa, 2013), while others find the exact opposite relation (Hanka, 1998; Cronqvist et al., 2009; Michaels, et al., 2019). Implication 4 rationalizes why high leverage can also help improve retention and suggests how to identify which firms are more likely to exhibit such a pattern.

Another lever through which firms can affect retention for a given level of total pay is by compensating (identical) workers differently. In particular, by ensuring that some "key" workers have a higher probability of staying through safer or higher compensation, the firm can ensure the retention of a critical mass of workers. In turn, this reduces contagious turnover, i.e., makes also the remaining workers less likely to leave. There is abundant anecdotal evidence of firms pursuing such strategies in practice. For example, in June 2021, the financial press extensively reported that Credit Suisse was targeting only a subset of its employees with compensation increases the prevent the mass exodus of its overall staff, concerned with being "the last man standing." We predict that the optimal compensation offer to these select "targeted" workers depends again on the firm's (relative) sensitivity to idiosyncratic and systematic risk (see Proposition 5):

Implication 5 *Overall retention for a given total level of compensation can be improved by offering (identical) workers different compensation structures. In particular, financially-constrained firms with high relative sensitivity to risk can improve the retention of select employees by offering them higher fixed (safer) compensation.*

While from a theoretical perspective, the optimality of asymmetric compensation for identical workers is particularly interesting, workers are rarely identical in practice. Accordingly, Implication 5 should be interpreted in the sense that even marginal differences among workers (which create a marginally higher preference to retain some workers) could lead to large differences in optimal compensation design.

7 Conclusion

The high quit rate at human capital intensive firms and the trillion dollars cost related to rehiring and training new workers makes improving retention a first-order priority for firms. While avoiding the departure of individual workers is certainly costly, a more substantial threat is that turnover is often contagious: The departure of skilled workers erodes firm performance, which might trigger other workers to leave as well. In this paper, we show that compensation design plays a crucial role in mitigating such collective turnover.

Our first main result shows the importance of offering dilutable (equity-based) pay to achieve cost-efficient retention in a multiple-worker setting. In particular, since workers forgo their deferred equity compensation when leaving, the remaining workers incur less dilution. Intuitively, while the size of the pie may decrease, the remaining workers' share of the pie increases, reducing their incentives to leave. Conversely, diluting workers' equity-pay is optimal when retention levels are high and the firm is doing well, as workers' on-the-job pay remains high despite dilution. Overall, dilution lowers the strategic complementarity in workers' decision to stay or leave, which reduces the probability of a worker run, while simultaneously lowering workers' rents. In practice, making compensation dilutable could be achieved via (deferred) equity stock option pools or bonus-sharing schemes.

Second, contagious turnover could also be prevented by offering fixed compensation that does not depend on the retention of other workers. The optimal mix between fixed and dilutable pay depends on the firms' sensitivity to risk relative to that of workers' outside options. Firms with lower sensitivity to systematic risk relative to workers' outside options will offer more dilutable (equity-based) relative to fixed pay. The general principle is that workers' on-the-job pay should match their outside options as closely as possible to achieve retention in a cost-efficient way. Thus, when workers' outside options are more sensitive to shocks than firm cash flows, firms must offer more equity-based pay to make on-the-job compensation more sensitive to shocks.

Our theory also provides implications for optimal security design since different choices of external financing might relax or tighten the resource constraints preventing cost-efficient retention. Interestingly, we obtain that higher leverage could lower the retention costs in firms with low exposure to systematic shocks. These predictions are reversed for firms with high exposure to systematic shocks relative to workers' outside options. Such firms are better off using credit lines or equity financing. These results can help explain why half of the empirical evidence on the relation between leverage and compensation finds a positive, while the other half finds a negative relation.

Finally, we show that minimizing collective turnover may call for compensating ex-ante identical workers differently to ensure that a critical mass of workers is always retained, in turn, reducing contagious turnover. In practice, this means that small differences among workers could lead to large differences in compensation design. This 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. In particular, firms with high sensitivity to risk relative to workers' outside options can improve retention of select workers by offering them safer compensation. Overall, our paper shows that mitigating collective turnover poses additional challenges relative to individual turnover, which require different solutions.

References

[1] Aldatmaz, Serdar, Paige Ouimet, Edward D Van Wesep, 2018, The option to quit: the effect of employee stock options on turnover, *Journal of Financial Economics* 127(1), 136–151.

[2] Agrawal, Ashwini K., and David A. Matsa, 2013, Labor unemployment risk and corporate financing decisions, *Journal of Financial Economics* 108(2), 449–470.

[3] Bergman, Nittai K., and Dirk Jenter, 2007, Employee sentiment and stock option compensation, *Journal of Financial Economics* 84(3), 667–712.

[4] Berk, Jonathan B., Richard Stanton, and Josef Zechner, 2010, Human capital, bankruptcy, and capital structure, *Journal of Finance* 65(3), 891-926.

[5] Breitling, Frank, Julia Dhar, Ruth Ebeling, and Deborah Lovich, 2021, 6 Strategies to Boost Retention Through the Great Resignation, *Harvard Business Review*, retrieved from https://hbr.org/2021/11/6-strategies-to-boost-retention-through-the-great-resignation

[6] Brown, Jennifer, and David Matsa, 2016, Boarding a sinking ship? An investigation of job applications to distressed firms, *Journal of Finance* 71(2), 507–550.

[7] Chemmanur, Thomas J., Yingmei Cheng, Tianming Zhang, 2013, Human capital, capital structure, and employee pay: an empirical analysis, *Journal of Financial Economics* 110(2), 478–502.

[8] Chen, Alvin, 2020, Firm performance pay as insurance against promotion risk, Working Paper, University of Washington.

[9] 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(1), 87–113.

[10] Cronqvist, Henrik, Fredrik Heyman, Mattias Nilsson, Helena Svaleryd, and Jonas Vlachos, 2009, Do entrenched managers pay their workers more?, *Journal of Finance* 64(1), 309–339.

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

[12] Döttling, Robin, Tomislav Ladika, and Enrico Perotti, 2019, Creating intangible capital, Working Paper, University of Amsterdam.

[13] 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(3), 545–561.

[14] Ferreira, Daniel, and Radoslawa Nikolowa, 2019, Chasing lemons: competition for talent under asymmetric information, Working Paper, London School of Economics and Queen Mary University.

[15] Fulghieri, Paolo, and David Dicks, 2019, Uncertainty and contracting: a theory of consensus and envy in organizations, Working Paper, University of North Carolina and Baylor University.

[16] Goldstein, Itay, and Ady Pauzner, 2005, Demand-deposit contracts and the probability of bank runs, *Journal of Finance* 60(3), 1293–1327.

[17] Grossman, Sanford. J., and Oliver. D. Hart, 1983, An analysis of the principal-agent problem, *Econometrica* 51, 7–45.

[18] Halac, Marina, Elliot Lipnowski, and Daniel Rappoport, 2021, Rank uncertainty in organizations, *American Economic Review* 111(3), 757–86.

[19] Hancock, Jullie I., David G. Allen, Frank A. Bosco, Karen R. McDaniel, Charles A. Pierce, 2013, Meta-Analytic Review of Employee Turnover as a Predictor of Firm Performance, *Journal of Management* 39(3), 573-603.

[20] Hand, John R. M., 2008, Give everyone a prize? Employee stock options in private venture-backed firms, *Journal of Business Venturing* 23(4), 385–404.

[21] Hanka, Gordon, 1998, Debt and the terms of employment, *Journal of Financial Economics* 48(3), 245–282.

[22] 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(1), 352–388.

[23] 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(3), 412–453.

[24] Hoberg, Gerard, and Gordon Phillips, 2016, Text-based network industries and endogenous product differentiation, *Journal of Political Economy* 124(5), 1423–1465.

[25] Hochberg, Yael V., and Laura Lindsey, 2010, Incentives, targeting and firm performance: an analysis of non-executive stock options, *Review of Financial Studies* 23(11), 4148–4186.

[26] Hoffmann, Florian, Roman Inderst, and Marcus Opp, 2021, The economics of deferral and clawback requirements, *Journal of Finance* (forthcoming).

[27] Holmström, Bengt, 1982, Moral hazard in teams, *Bell Journal of Economics* 13(2), 324–340.

[28] Innes, Robert D., 1990, Limited liability and incentive contracting with ex-ante action choices, *Journal of Economic Theory* 52(1), 45–67.

[29] Jochem, Torsten, Tomislav Ladika, and Zacharias Sautner, 2018, The retention effects of unvested equity: evidence from accelerated option vesting, *Review of Financial Studies* 31(11), 4142–4186.

[30] Lazear, Edward, 2004. Output-based pay: incentives, retention or sorting?, in Solomon W. Polachek (ed.) Accounting for Worker Well-Being (Research in Labor Economics, Volume 23) Emerald Group Publishing Limited, 1 - 25.

[31] Michaels, Ryan, T. Beau Page, Toni M. Whited, 2019, Labor and capital dynamics under financing frictions, *Review of Finance* 23(2), 279–323.

[32] 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.

[33] Murphy, Kevin J., 2003, Stock-based pay in new economy firms, *Journal of Accounting and Economics* 34 (1-3), 129-147.

[34] Nachman, David C., Noe, Thomas H., 1994, Optimal design of securities under asymmetric information, *Review of Financial Studies* 7(1), 1–44.

[35] Oyer, Paul, 2004, Why do firms use incentives that have no incentive effects?, *Journal of Finance* 59(4), 1619–1649

[36] Perotti, Enrico C., and Kathryn E. Spier, 1993, Capital structure as a bargaining tool: the role of leverage in contract renegotiation, *American Economic Review* 85(5), 1131–1141.

[37] Rajan, Raghuram G., 2012, Presidential address: the corporation in finance, *Journal of Finance* 67(4), 1173–1217.

[38] Titman, Sheridan, 1984, The effect of capital structure on a firm's liquidation decision, *Journal of Financial Economics* 13(1), 137–151.

[39] Van Wesep, Edward D., and Brian Waters, 2021, Bonus season: a theory of periodic labor markets and coordinated bonuses, *Management Science,* forthcoming.

[40] Winter, Eyal, 2004, Incentives and discrimination, *American Economic Review* 94 (3), 764–73.

Appendix A Proofs

Proof of Proposition 1. The proof is standard and, therefore, omitted. **Q.E.D.**

Proof of Lemma 1. (i) The workers' payoff function satisfies all standard assumptions in the global games literature. Specifically, $w(n)+p(n,\varepsilon)\Delta w(n)-\underline{w}(\varepsilon)$ is increasing in n and ε; there is a cutoff ε^* that satisfies (6) and there are shock realizations $\varepsilon', \varepsilon'' \in \mathbb{R}$ and $\xi > 0$, such that $w+p(n,\varepsilon)\Delta w-\underline{w}(\varepsilon) > \xi$ for all $n \in [1,N]$ and $\varepsilon \geq \varepsilon'$ and $w+p(n,\varepsilon)\Delta w-\underline{w}(\varepsilon) \leq -\xi$ for all $n \in [1,N]$ and $\varepsilon \leq \varepsilon''$. Finally, $\int_{z=-\infty}^{\infty} zf(z)\,dz$ is well defined. Hence, for any $\delta > 0$, there is a $\overline{\sigma} > 0$, such that for all $\sigma \leq \overline{\sigma}$, there is a cutoff equilibrium in which the workers run if they observe $\widetilde{\varepsilon} < \varepsilon^* - \delta$ and stay if $\widetilde{\varepsilon} > \varepsilon^* + \delta$ (see Appendix B in Morris and Shin, 2003).

(ii) This case is the mirror image of case (i). Let $\underline{n} = N - n$ be the number of workers that *leave*. It holds that $\underline{w}(\varepsilon) - w(\underline{n}) - p(\underline{n},\varepsilon)\Delta w(\underline{n})$ is increasing in $\underline{n}$ and ε; there is a cutoff ε^* that satisfies (6); and there are shock realizations $\varepsilon', \varepsilon'' \in \mathbb{R}$ and $\xi > 0$, such that $\underline{w}(\varepsilon) - w(\underline{n}) - p(n,\varepsilon)\Delta w(n) > \xi$ for all $\underline{n} \in [1,N]$ and $\varepsilon \geq \varepsilon''$ and $\underline{w}(\varepsilon) - w(\underline{n}) - p(\underline{n},\varepsilon)\Delta w(\underline{n}) \leq -\xi$ for all $\underline{n} \in [1,N]$ and $\varepsilon \leq \varepsilon'$. Thus, by analogous arguments to Morris and Shin (2003), for any $\delta > 0$, there is a $\overline{\sigma} > 0$, such that for all $\sigma \leq \overline{\sigma}$, there is a cutoff equilibrium in which the workers run if they observe $\widetilde{\varepsilon} > \varepsilon^* + \delta$ and stay if $\widetilde{\varepsilon} < \varepsilon^* - \delta$. **Q.E.D.**

Proof of Proposition 2. The contract characterized in the proposition satisfies condition (9) for all (ε, n). Therefore, it achieves full retention and full rent extraction. **Q.E.D.**

Proof of Proposition 3. The contract characterized in the proposition satisfies condition (9) for all (ε, n). Therefore, it achieves full retention and full rent extraction. **Q.E.D.**

Proof of Proposition 4. We start by introducing some notation. Let $g(\varepsilon|\widetilde{\varepsilon}_i)$ denote the conditional distribution of ε given a signal realization $\widetilde{\varepsilon}_i$ and $h(\widetilde{\varepsilon}_i|\varepsilon)$ denote the conditional distribution of $\widetilde{\varepsilon}_i$ given a shock realization ε. Since the joint distribution of ε and $\widetilde{\varepsilon}_i$ satisfies the monotone likelihood ratio property (MLRP), so do $g(\varepsilon|\widetilde{\varepsilon}_i)$ and $h(\widetilde{\varepsilon}_i|\varepsilon)$. A useful property of MLRP is that it implies first order stochastic dominance. In what follows, denote the probability of observing signal $\widetilde{\varepsilon}_i$ with $h(\widetilde{\varepsilon}_i) \equiv \int_{-\infty}^{\infty} h(\widetilde{\varepsilon}_i|\varepsilon)\, g(\varepsilon)\, d\varepsilon$ and observe that the number of retained workers, n, is determined by all workers' signal realizations (Lemma 1). Finally, denote $W(\varepsilon, n) = w(n) + p(\varepsilon, n)\Delta w(n)$.

Part (i) We, now, prove that for any n, the payment to the worker in the low cash flow state is set to the highest value allowed by the resource constraint $w(n) \leq \frac{x}{n}$ and the monotonicity constraints, $\boldsymbol{\beta}^{\mathbf{n}} \geq 0$. The proof is by contradiction. Suppose to a contradiction that for some $n = \widetilde{n}$, it holds $w(\widetilde{n}) < \frac{x}{\widetilde{n}}$. Construct an alternative contract $\{\widetilde{w}(n), \Delta\widetilde{w}(n)\}$ that is identical to $\{w(n), \Delta w(n)\}$ for all retention levels $n \neq \widetilde{n}$ but stipulates that $\widetilde{w}(\widetilde{n}) = w(\widetilde{n}) + \zeta$ and $\Delta\widetilde{w}(\widetilde{n}) = \Delta w(\widetilde{n}) - \xi$. Assume for now that $\Delta w(\widetilde{n}) > 0$ and $W(\varepsilon, \widetilde{n}+1) > W(\varepsilon, \widetilde{n}) > W(\varepsilon, \widetilde{n}-1)$ for all ε so that this perturbation is feasible without violating any constraints.

Let ξ and ζ be such that the cutoff signal, ε^*, for the new contract is the same as for the old one. Since the worker must be indifferent between staying and leaving at ε^* with both the original and perturbed contracts, it must hold

$$
\begin{aligned}
0 &= \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}} \left[w(n) + p(\varepsilon, n) \Delta w(n) - \underline{w}(\varepsilon) | \varepsilon, \varepsilon^* \right] dG(\varepsilon|\varepsilon^*) && \text{(A.1)} \\
&= \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}} \left[\widetilde{w}(n) + p(\varepsilon, n) \Delta\widetilde{w}(n) - \underline{w}(\varepsilon) | \varepsilon, \varepsilon^* \right] dG(\varepsilon|\varepsilon^*) && \text{(A.2)} \\
&= \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}} \left[w(n) + p(\varepsilon, n) \Delta w(n) - \underline{w}(\varepsilon) | \varepsilon, \varepsilon^* \right] dG(\varepsilon|\varepsilon^*) && \text{(A.3)} \\
&\quad + \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}} \left[(\zeta - p(\varepsilon, \widetilde{n}) \xi) \mathbf{1}_{\widetilde{n}} | \varepsilon, \varepsilon^* \right] dG(\varepsilon|\varepsilon^*), && \text{(A.4)}
\end{aligned}
$$

where $\mathbf{1}_{\widetilde{n}}$ is an indicator function equal to one if $\widetilde{n}$ workers stay (where we assume that the worker stays when indifferent). Note that for any $\widetilde{\varepsilon}_i \geq \varepsilon^*$ it holds that

$$
\mathrm{E}_{\widetilde{\varepsilon}_{-i}} \left[\mathbf{1}_{\widetilde{n}} | \varepsilon, \varepsilon^* \right] = \begin{pmatrix} N-1 \\ \widetilde{n}-1 \end{pmatrix} (1 - H(\varepsilon^*|\varepsilon))^{\widetilde{n}-1} H(\varepsilon^*|\varepsilon)^{N-\widetilde{n}}. \tag{A.5}
$$

Using from expressions (A.1) and (A.3) that $0 = \int_{-\infty}^{\infty} \mathrm{E}_{\widetilde{\varepsilon}_{-i}} \left[(\zeta - p(\varepsilon, \widetilde{n}) \xi) \mathbf{1}_{\widetilde{n}} | \varepsilon, \varepsilon^* \right] dG(\varepsilon|\varepsilon^*)$ and using that $p(\varepsilon, \widetilde{n})$ increases in ε, we obtain that $\zeta - p(\varepsilon, \widetilde{n}) \xi$ crosses ε once from above for some shock realization $\widehat{\varepsilon} \in [\underline{\varepsilon}, \overline{\varepsilon}]$ for which it holds that $\zeta = p(\varepsilon, \widetilde{n}) \xi$. From this, we obtain that for any signal $\widetilde{\varepsilon}_i > \varepsilon^*$, it holds

$$
\begin{aligned}
&\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}} \left[(\zeta - p(\varepsilon, \widetilde{n}) \xi) \mathbf{1}_{\widetilde{n}} | \varepsilon, \widetilde{\varepsilon}_i \right] dG(\varepsilon|\widetilde{\varepsilon}_i) \\
= &\int_{\underline{\varepsilon}}^{\widehat{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}} \underbrace{\left[(\zeta - p(\varepsilon, \widetilde{n}) \xi) \mathbf{1}_{\widetilde{n}} | \varepsilon, \widetilde{\varepsilon}_i \right]}_{\geq 0} \frac{g(\varepsilon|\widetilde{\varepsilon}_i)}{g(\varepsilon|\varepsilon^*)} g(\varepsilon|\varepsilon^*) \, d\varepsilon \\
&+ \int_{\widehat{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}} \underbrace{\left[(\zeta - p(\varepsilon, \widetilde{n}) \xi) \mathbf{1}_{\widetilde{n}} | \varepsilon, \widetilde{\varepsilon}_i \right]}_{\leq 0} \frac{g(\varepsilon|\widetilde{\varepsilon}_i)}{g(\varepsilon|\varepsilon^*)} g(\varepsilon|\varepsilon^*) \, d\varepsilon \\
< &\frac{g(\widehat{\varepsilon}|\widetilde{\varepsilon}_i)}{g(\widehat{\varepsilon}|\varepsilon^*)} \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}} \left[(\zeta - p(\varepsilon, \widetilde{n}) \xi) \mathbf{1}_{\widetilde{n}} | \varepsilon, \varepsilon^* \right] dG(\varepsilon|\varepsilon^*) = 0, && \text{(A.6)}
\end{aligned}
$$

where we use that $\frac{g(\varepsilon|\widetilde{\varepsilon}_i)}{g(\varepsilon|\varepsilon^*)}$ is increasing in ε by MLRP and that $\mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\mathbf{1}_{\widetilde{n}}|\varepsilon,\varepsilon^*\right] = \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\mathbf{1}_{\widetilde{n}}|\varepsilon,\widetilde{\varepsilon}_i\right]$ for any $\widetilde{\varepsilon}_i \geq \varepsilon^*$ (see expression (A.5)).

From expression (A.6), it follows that the difference in ex ante expected payoffs between the two contracts is

$$\begin{aligned}
&\int_{\varepsilon^*}^{\infty}\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\widetilde{w}\left(n\right)+p\left(\varepsilon,n\right)\Delta\widetilde{w}\left(n\right)|\varepsilon,\widetilde{\varepsilon}_i\right]dG\left(\varepsilon|\widetilde{\varepsilon}_i\right)dH\left(\widetilde{\varepsilon}_i\right)\\
&-\int_{\varepsilon^*}^{\infty}\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[w\left(n\right)+p\left(\varepsilon,n\right)\Delta w\left(n\right)|\varepsilon,\widetilde{\varepsilon}_i\right]dG\left(\varepsilon|\widetilde{\varepsilon}_i\right)dH\left(\widetilde{\varepsilon}_i\right)\\
=&\int_{\varepsilon^*}^{\infty}\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\left(\zeta-p\left(\varepsilon,\widetilde{n}\right)\xi\right)\mathbf{1}_{\widetilde{n}}|\varepsilon,\widetilde{\varepsilon}_i\right]dG\left(\varepsilon|\widetilde{\varepsilon}_i\right)dH\left(\widetilde{\varepsilon}_i\right)<0.
\end{aligned}$$

Thus, contract $\{\widetilde{w}(n), \Delta\widetilde{w}(n)\}$ implements the same cutoff ε^* while leaving less rent to the workers that observe a signal $\widetilde{\varepsilon}_i > \varepsilon^*$ and stay. This gives us the desired contradiction.

We assumed above that $\Delta w(\widetilde{n}) > 0$ and $W(\varepsilon, \widetilde{n}) > W(\varepsilon, \widetilde{n}-1)$ with the original contract. It is worth remarking that if $W(\varepsilon, \widetilde{n}) = W(\varepsilon, \widetilde{n}-1)$ for the original contract, then we can perturb the contract for a retention level of $n-1$ and possibly also lower retention levels to make the perturbation of $\{w(\widetilde{n}), \Delta w(\widetilde{n})\}$ possible. Since all these perturbations involve increasing w while decreasing Δw at the respective retention level, the same argument as above applies (see Lemma A.1 following this proof for details).

Finally, consider the case in which $\Delta w(\widetilde{n}) = 0$, so that reducing $\Delta w(\widetilde{n})$ is not feasible. In this case, the increase in $w(n)$ is offset by reducing another $\Delta w(n') > 0$ for $n' > \widetilde{n}$. To see that there is at least one such n' observe that by monotonicity in n, all contracts for $n \leq \widetilde{n}$ must be yielding an expected compensation of less than $\mathrm{E}\left[W(\varepsilon, \widetilde{n})\right] \leq \frac{x}{\widetilde{n}} < \underline{w}$ (the first inequality follows from $\Delta w(\widetilde{n}) = 0$ and $w(\widetilde{n}) \leq \frac{x}{\widetilde{n}}$). Thus, if the workers are to break even for some interior ε^*, there must be at least one $\Delta w(n') > 0$ for $n' > \widetilde{n}$.

Consider a perturbation of $\widetilde{w}(\widetilde{n}) = w(\widetilde{n}) + \zeta$ and $\Delta w(n') = \Delta w(n') - \xi$ such that the cutoff ε^* remains unchanged — i.e., $\zeta = p(\varepsilon^*, n')\,\xi$. Since the workers must be indifferent between the two contracts, it holds that:

$$0=\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[p\left(\varepsilon^*,n'\right)\xi\mathbf{1}_{\widetilde{n}}-\xi p\left(\varepsilon,n'\right)\mathbf{1}_{n'}|\varepsilon,\varepsilon^*\right]dG\left(\varepsilon|\varepsilon^*\right) \tag{A.7}$$

In Lemma A.2 following this proof, we show that the term in square brackets of expression (A.7) crosses zero in ε at most once from above at some $\widehat{\varepsilon} \in [\underline{\varepsilon}, \overline{\varepsilon}]$. Thus, we can follow the same steps as in the case when $\Delta w(\widetilde{n}) > 0$ to show that the deviation is profitable for the firm. **Q.E.D.**

Part (ii). We prove that for any n, $w(n)$ is minimized subject to the limited liability constraint $w(n) \geq 0$ and the monotonicity constraints, $\boldsymbol{\beta}^{\mathbf{n}} \geq 0$. The proof is a straightforward modification of the proof of part (i) and is again by contradiction. Suppose that a contract $\{w(n), \Delta w(n)\}$ were optimal for which $w(n) > 0$ for at least one $n = \widetilde{n}$. Construct an alternative contract $\{\widetilde{w}(n), \Delta\widetilde{w}(n)\}$ that is the same as $\{w(n), \Delta w(n)\}$ for all $n \neq \widetilde{n}$ but $\widetilde{w}(\widetilde{n}) = w(\widetilde{n}) - \zeta$ and $\Delta\widetilde{w}(\widetilde{n}) = \Delta w(\widetilde{n}) + \xi$ for $n = \widetilde{n}$. Assume for now that $\Delta w(\widetilde{n}) < \frac{\Delta x}{\widetilde{n}}$ and $W(\varepsilon, \widetilde{n}+1) > W(\varepsilon, \widetilde{n}) > W(\varepsilon, \widetilde{n}-1)$ for all ε so that this perturbation is feasible without violating any constraints.

Choose, now, ζ and ξ, so that the cutoff signal above which the workers leave the firm is the same with both contracts, i.e., ε^*. Since the workers need to be indifferent between staying and leaving at ε^* it holds

$$\begin{aligned}
0 &= \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[w(n) + p(\varepsilon, n)\Delta w(n) - \underline{w}(\varepsilon)\,|\varepsilon, \varepsilon^*\right] dG(\varepsilon|\varepsilon^*) \\
&= \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\widetilde{w}(n) + p(\varepsilon, n)\Delta\widetilde{w}(n) - \underline{w}(\varepsilon)\,|\varepsilon, \varepsilon^*\right] dG(\varepsilon|\varepsilon^*) \qquad \text{(A.8)} \\
&= \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[w(n) + p(\varepsilon, n)\Delta w(n) - \underline{w}(\varepsilon)\,|\varepsilon, \varepsilon^*\right] dG(\varepsilon|\varepsilon^*) \\
&\quad + \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\left(p(\varepsilon, \widetilde{n})\xi - \zeta\right)\mathbf{1}_{\widetilde{n}}|\varepsilon, \varepsilon^*\right] dG(\varepsilon|\varepsilon^*),
\end{aligned}$$

implying that $0 = \int_{-\infty}^{\infty} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\left(p(\varepsilon, \widetilde{n})\xi - \zeta\right)\mathbf{1}_{\widetilde{n}}|\varepsilon, \varepsilon^*\right] dG(\varepsilon|\varepsilon^*)$. Using that $p(\varepsilon, \widetilde{n})\xi - \zeta$ increases in ε, it follows that there is a shock realization $\widehat{\varepsilon}$ for which it holds that $p(\varepsilon, \widetilde{n})\xi = \zeta$. From this, we obtain that for any signal $\widetilde{\varepsilon}_i \leq \varepsilon^*$, it holds

$$\begin{aligned}
&\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\left(p(\varepsilon, \widetilde{n})\xi - \zeta\right)\mathbf{1}_{\widetilde{n}}|\varepsilon, \widetilde{\varepsilon}_i\right] dG(\varepsilon|\widetilde{\varepsilon}_i) \\
= &\int_{\underline{\varepsilon}}^{\widehat{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\underbrace{\left[\left(p(\varepsilon, \widetilde{n})\xi - \zeta\right)\mathbf{1}_{\widetilde{n}}|\varepsilon, \widetilde{\varepsilon}_i\right]}_{\leq 0} \frac{g(\varepsilon|\widetilde{\varepsilon}_i)}{g(\varepsilon|\varepsilon^*)} g(\varepsilon|\varepsilon^*)\, d\varepsilon \\
&+ \int_{\widehat{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\underbrace{\left[\left(p(\varepsilon, \widetilde{n})\xi - \zeta\right)\mathbf{1}_{\widetilde{n}}|\varepsilon, \widetilde{\varepsilon}_i\right]}_{\geq 0} \frac{g(\varepsilon|\widetilde{\varepsilon}_i)}{g(\varepsilon|\varepsilon^*)} g(\varepsilon|\varepsilon^*)\, d\varepsilon \\
< &\ \frac{g(\widehat{\varepsilon}|\widetilde{\varepsilon}_i)}{g(\widehat{\varepsilon}|\varepsilon^*)} \int_{\underline{\varepsilon}}^{\widetilde{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\left(p(\varepsilon, \widetilde{n})\xi - \zeta\right)\mathbf{1}_{\widetilde{n}}|\varepsilon, \widetilde{\varepsilon}_i\right] dG(\varepsilon|\varepsilon^*) = 0, \qquad \text{(A.9)}
\end{aligned}$$

where we use that $\frac{g(\varepsilon|\widetilde{\varepsilon}_i)}{g(\varepsilon|\varepsilon^*)}$ is decreasing in ε by MLRP (as $\widetilde{\varepsilon}_i \leq \varepsilon^*$).

From expression (A.9), it follows that the difference in ex ante expected payoffs between contract $\{\widetilde{w}(n), \Delta\widetilde{w}(n)\}$ and $\{w(n), \Delta w(n)\}$ is

$$\int_{-\infty}^{\varepsilon^*}\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\left(p(\varepsilon,\widetilde{n})\xi-\zeta\right)\mathbf{1}_{\widetilde{n}}|\varepsilon,\widetilde{\varepsilon}_i\right] g(\varepsilon|\widetilde{\varepsilon}_i)\, d\varepsilon dH(\widetilde{\varepsilon}_i) < 0.$$

Thus, contract $\{\widetilde{w}(n), \Delta\widetilde{w}(n)\}$ could implement the same cutoff ε^* while leaving less rent to the workers that observe a signal $\widetilde{\varepsilon}_i < \varepsilon^*$ and stay. This gives us the desired contradiction.

As in the proof of part (i), it is worth remarking that if one of the monotonicity constraints in n is binding in $\widetilde{n}$, we could perturb the contracts adjacent to $\widetilde{n}$ to make the perturbation of $\{w(\widetilde{n}), \Delta w(\widetilde{n})\}$ possible. Since all these perturbations involve decreasing $w(n)$ while increasing $\Delta w(n)$ at the respective retention level, the same argument as above applies.

Finally, we consider the case in which $\Delta w(\widetilde{n}) = \frac{\Delta x}{\widetilde{n}}$, which requires that the decrease in $w(\widetilde{n})$ is offset by increasing another $\Delta w(n') > 0$ for $n' > n$. Consider a perturbation of $\widetilde{w}(\widetilde{n}) = w(\widetilde{n}) - \zeta$ and $\Delta w(n') = \Delta w(n') + \xi$ such that the ε^* remains unchanged — i.e., $\zeta = p(\varepsilon^*, n')\xi$. Since the workers must be indifferent between the two contracts, it holds that:

$$0 = \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\xi p(\varepsilon, n')\mathbf{1}_{n'} - p(\varepsilon^*, n')\xi\mathbf{1}_{\widetilde{n}}|\varepsilon,\varepsilon^*\right] dG(\varepsilon|\varepsilon^*) \tag{A.10}$$

We can, now, modify Lemma A.2 to show that the term in square brackets of expression (A.10) crosses zero in ε at most once from below at some $\widehat{\varepsilon} \in [\underline{\varepsilon}, \overline{\varepsilon}]$. Thus, we can follow the same steps as in the case in which $\Delta w(\widetilde{n}) > 0$ to show that the deviation is profitable for the firm. **Q.E.D.**

Lemma A.1 *Suppose that $W(\varepsilon, n) \le W(\varepsilon, n-1)$ binds for some ε and n. Then this monotonicity constraint can be relaxed by increasing $w(n)$ and decreasing $\Delta w(n)$ or increasing $w(n+1)$ and decreasing $\Delta w(n+1)$, while keeping the same cutoff ε^*.*

Proof of Lemma A.1. Suppose that there is a binding monotonicity constraint $W(\varepsilon, n') = W(\varepsilon, n'-1)$. For use below, observe that for any perturbation at some $n = n'$ for which the cutoff ε^* remains the same, it must hold that

$$w(n') + \zeta + p(\varepsilon^*, n')(\Delta w(n') - \xi) = w(n') + p(\varepsilon^*, n')\Delta w(n').$$

Hence, it must be that $\zeta = p(\varepsilon^*, n')\xi$. We can relax a binding monotonicity constraint in n in one of three ways.

(i) If $W(\varepsilon, n) = W(\varepsilon, n'-1)$ at some $\varepsilon' < \varepsilon^*$, then by constructing a deviation for $n = n'$

(as outlined above), we obtain

$$\begin{aligned}\widetilde{w}(n') + p(\varepsilon', n')\Delta\widetilde{w}(n') &= w(n') + \zeta + p(\varepsilon', n')(\Delta w(n') - \xi) \\ &= w(n') + p(\varepsilon', n')\Delta w(n'-1) + \underbrace{(p(\varepsilon^*, n') - p(\varepsilon', n'))\xi}_{>0} \\ &> w(n'-1) + p(\varepsilon', n')\Delta w(n'-1)\end{aligned}$$

so the monotonicity constraint at n' is relaxed at ε'.

(ii) Suppose that the monotonicity constraint at n' binds at some $\varepsilon'' > \varepsilon^*$. Then by constructing a deviation for $n'-1$ (as outlined above), we obtain

$$\begin{aligned}&\widetilde{w}(n'-1) + p(\varepsilon'', n'-1)\Delta\widetilde{w}(n'-1) \\ = \;& w(n'-1) + p(\varepsilon'', n'-1)\Delta w(n'-1) + \underbrace{(p(\varepsilon^*, n'-1) - p(\varepsilon'', n'-1))\xi}_{<0} \\ < \;& w(n') + p(\varepsilon'', n')\Delta w(n')\end{aligned}$$

so the monotonicity constraint at n is relaxed at ε''.

(iii) Suppose that the monotonicity constraint at n' binds at ε^* or all ε. Then, we can perform a simultaneous perturbation of the contracts for n' and $n'-1$ as stated in cases (i) and (ii) that leaves ε^* unchanged without violating monotonicity.

It follows that, if any of the perturbations from Proposition 4 is prevented by an adjacent monotonicity constraint, then we can potentially perform the same perturbations as in the proof of Proposition 4 to relax these adjacent monotonicity constraints. Since ε^* is kept constant for all contract perturbations, and for all these perturbations, we have increased $w(n')$ and decreased $\Delta w(n')$, the proof of Proposition 4 applies almost unchanged. **Q.E.D.**

Lemma A.2 *If $n' > \widetilde{n}$, it holds that*

$$\mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[p(\varepsilon^*, n')\xi\mathbf{1}_{\widetilde{n}}|\varepsilon, \varepsilon^*\right] - \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\xi p(\varepsilon, n')\mathbf{1}_{n'}|\varepsilon, \varepsilon^*\right] \tag{A.11}$$

crosses zero in ε at most once from above.

Proof of Lemma A.2. The term in square brackets of expression (A.11) is equal to

$$
\begin{aligned}
& p\left(\varepsilon^*, n'\right) \xi \begin{pmatrix} N-1 \\ \widetilde{n}-1 \end{pmatrix} \left(1-H\left(\varepsilon^*|\varepsilon\right)\right)^{\widetilde{n}-1} H\left(\varepsilon^*|\varepsilon\right)^{N-\widetilde{n}} \\
& -p\left(\varepsilon, n'\right) \xi \begin{pmatrix} N-1 \\ n'-1 \end{pmatrix} \left(1-H\left(\varepsilon^*|\varepsilon\right)\right)^{n'-1} H\left(\varepsilon^*|\varepsilon\right)^{N-n'} \\
= & \left(p\left(\varepsilon^*, n'\right) \begin{pmatrix} N-1 \\ \widetilde{n}-1 \end{pmatrix} H\left(\varepsilon^*|\varepsilon\right)^{n'-\widetilde{n}} - p\left(\varepsilon, n'\right) \begin{pmatrix} N-1 \\ n'-1 \end{pmatrix} \left(1-H\left(\varepsilon^*|\varepsilon\right)\right)^{n'-\widetilde{n}}\right) \\
& \times \xi \left(1-H\left(\varepsilon^*|\varepsilon\right)\right)^{\widetilde{n}-1} H\left(\varepsilon^*|\varepsilon\right)^{N-n'}
\end{aligned}
$$

Suppose that this term is zero at some $\widehat{\varepsilon}$. Since expression (A.7) is zero, there must be at least one such $\widehat{\varepsilon}$:

$$
0 = p\left(\varepsilon^*, n'\right) \begin{pmatrix} N-1 \\ \widetilde{n}-1 \end{pmatrix} H\left(\varepsilon^*|\varepsilon\right)^{n'-\widetilde{n}} - p\left(\varepsilon, n'\right) \begin{pmatrix} N-1 \\ n'-1 \end{pmatrix} \left(1-H\left(\varepsilon^*|\varepsilon\right)\right)^{n'-\widetilde{n}}. \tag{A.12}
$$

By first-order stochastic dominance, $H\left(\varepsilon^*|\varepsilon\right)$ decreases in ε. Hence, expression (A.12) strictly decreases in ε, and the crossing at zero must be from above. Furthermore, since expression (A.12) is continuous in ε, there can also be at most one such crossing. **Q.E.D.**

Proof of Proposition 5. Suppose that there is an equilibrium in which worker i only stays if she observes a signal $\widetilde{\varepsilon}_i > \varepsilon_1^*$ while workers $j \neq i$ only if they observe a signal $\widetilde{\varepsilon}_j > \varepsilon_2^*$, where $\varepsilon_1^* < \varepsilon_2^*$. We show that in any such equilibrium, the firm will maximize $w_1(\mathbf{n})$ subject to the resource constraint $\frac{x}{n}$ and the monotonicity constraint $\boldsymbol{\beta}^{\mathbf{n}} \geq 0$ for all $\mathbf{n}$. Without loss of generality, we assume that the two groups of workers are of the same size.

The proof is by contradiction. Suppose to a contradiction that $w_1(\mathbf{n}) < \frac{x}{n_1}$ while $w_2(\mathbf{n}) > 0$ for some $\mathbf{n}$. Consider offering the two groups of workers alternative contracts for which the cutoffs ε_2^* and ε_1^* are the same as in the original contract. Specifically, for workers j, construct an alternative contract $\{\widetilde{w}_2(\mathbf{n}), \Delta\widetilde{w}_2(\mathbf{n})\}$ with $\widetilde{w}_2(\mathbf{n}) = w_2(\mathbf{n})$ for all $\mathbf{n} \neq \widetilde{\mathbf{n}}$ but $\widetilde{w}_2(\mathbf{n}) = w_2(\mathbf{n}) - \zeta$ and $\Delta\widetilde{w}_2(\widetilde{\mathbf{n}}) = \Delta w_2(\widetilde{\mathbf{n}}) + \xi$ for $\mathbf{n} = \widetilde{\mathbf{n}}$, where ζ and ξ are positive, but small, and it continues to hold that $\widetilde{w}_2(\mathbf{n}) + p(\varepsilon, \mathbf{n})\Delta\widetilde{w}_2(\mathbf{n})$ is non-decreasing in n. Furthermore, ξ is chosen such that ε_2^* is unchanged. Similarly, for workers i, construct a contract that is identical to the original one except that $\widetilde{w}_2(\mathbf{n}) = w_2(\mathbf{n}) - \zeta$ and $\Delta\widetilde{w}_2(\widetilde{\mathbf{n}}) = \Delta w_2(\widetilde{\mathbf{n}}) + \widetilde{\xi}$ for $\mathbf{n} = \widetilde{\mathbf{n}}$, where $\widetilde{\xi}$ is chosen such that ε_1^* is unchanged (the cases in which one of the feasibility constraints prevents such perturbations can be analyzed similarly to Proposition 4, respectively).

By the same arguments as in Proposition 4 (see expressions (A.6) and (A.9)), we obtain that

$$
\begin{aligned}
0 &= \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\varepsilon_{-i}}\left[\left(\zeta - p(\varepsilon,\mathbf{n})\widetilde{\xi}\right)\mathbf{1}_{\tilde{\mathbf{n}}}|\varepsilon,\varepsilon_1^*\right] dG(\varepsilon|\varepsilon_1^*) \\
&= \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\varepsilon_{-j}}\left[\left(\zeta - p(\varepsilon,\mathbf{n})\xi\right)\mathbf{1}_{\tilde{\mathbf{n}}}|\varepsilon,\varepsilon_2^*\right] dG(\varepsilon|\varepsilon_2^*) \qquad \text{(A.13)}
\end{aligned}
$$

and from MLRP that

$$
\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\left(\zeta - p(\varepsilon,\tilde{\mathbf{n}})\widetilde{\xi}\right)\mathbf{1}_{\tilde{\mathbf{n}}}|\varepsilon,\varepsilon_1^*\right] dG(\varepsilon|\widetilde{\varepsilon}_i)\begin{cases} <0 & \text{if } \widetilde{\varepsilon}_i > \varepsilon_1^* \\ >0 & \text{if } \widetilde{\varepsilon}_i < \varepsilon_1^* \end{cases}. \qquad \text{(A.14)}
$$

Since $\int_{-\infty}^{\infty} \mathrm{E}_{\varepsilon_{-j}}\left[\left(\zeta - p(\varepsilon,\mathbf{n})\xi\right)\mathbf{1}_{\tilde{\mathbf{n}}}|\varepsilon,\widetilde{\varepsilon}_j\right] dG(\varepsilon|\widetilde{\varepsilon}_j)$ is decreasing in $\widetilde{\varepsilon}_j$ and $\varepsilon_1^* < \varepsilon_2^*$, we further have from (A.13) that

$$
\begin{aligned}
0 &= \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\varepsilon_{-i}}\left[\left(\zeta - p(\varepsilon,\mathbf{n})\widetilde{\xi}\right)\mathbf{1}_{\tilde{\mathbf{n}}}|\varepsilon,\varepsilon_1^*\right] dG(\varepsilon|\varepsilon_1^*) \qquad \text{(A.15)} \\
&< \int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\varepsilon_{-j}}\left[\left(\zeta - p(\varepsilon,\mathbf{n})\xi\right)\mathbf{1}_{\tilde{\mathbf{n}}}|\varepsilon,\varepsilon_2^*\right] dG(\varepsilon|\varepsilon_1^*). \qquad \text{(A.16)}
\end{aligned}
$$

Since $\varepsilon_1^* < \varepsilon_2^*$, we can further replace $\mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\mathbf{1}_{\tilde{\mathbf{n}}}|\varepsilon,\varepsilon_1^*\right]$ with $\mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\mathbf{1}_{\tilde{\mathbf{n}}}|\varepsilon,\varepsilon_2^*\right]$ in line (A.15). Using this, and noting that all signals are drawn independently from the same distribution (i.e., for the same signal realizations, i and j have the same expectations), we obtain from (A.15) and (A.16) that $\widetilde{\xi} > \xi$.

Altogether, the change in the firm's expected payoffs from changing workers i's and j's contracts is

$$
\begin{aligned}
& -\int_{\varepsilon_1^*}^{\infty}\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\left(\zeta - p(\varepsilon,\mathbf{n})\widetilde{\xi}\right)\mathbf{1}_{\tilde{\mathbf{n}}}|\varepsilon,\widetilde{\varepsilon}_i\right] dG(\varepsilon|\widetilde{\varepsilon}_i)\, dH(\widetilde{\varepsilon}_i) \\
& +\int_{\varepsilon_2^*}^{\infty}\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-j}}\left[\left(\zeta - p(\varepsilon,\mathbf{n})\xi\right)\mathbf{1}_{\tilde{\mathbf{n}}}|\varepsilon,\widetilde{\varepsilon}_j\right] dG(\varepsilon|\widetilde{\varepsilon}_j)\, dH(\widetilde{\varepsilon}_j) \\
> & -\int_{\varepsilon_2^*}^{\infty}\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\left(\zeta - p(\varepsilon,\mathbf{n})\widetilde{\xi}\right)\mathbf{1}_{\tilde{\mathbf{n}}}|\varepsilon,\widetilde{\varepsilon}_i\right] dG(\varepsilon|\widetilde{\varepsilon}_i)\, dH(\widetilde{\varepsilon}_i) \\
& +\int_{\varepsilon_2^*}^{\infty}\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-j}}\left[\left(\zeta - p(\varepsilon,\mathbf{n})\xi\right)\mathbf{1}_{\tilde{\mathbf{n}}}|\varepsilon,\widetilde{\varepsilon}_j\right] dG(\varepsilon|\widetilde{\varepsilon}_j)\, dH(\widetilde{\varepsilon}_j) > 0
\end{aligned}
$$

where we use that $-\int_{\varepsilon_1^*}^{\varepsilon_2^*}\int_{\underline{\varepsilon}}^{\overline{\varepsilon}} \mathrm{E}_{\widetilde{\varepsilon}_{-i}}\left[\left(\zeta - p(\varepsilon,\mathbf{n})\widetilde{\xi}\right)\mathbf{1}_{\tilde{\mathbf{n}}}|\varepsilon,\widetilde{\varepsilon}_i\right] dG(\varepsilon|\widetilde{\varepsilon}_i) > 0$ (see expression (A.14)) to obtain the first inequality; the second inequality follows from $\widetilde{\xi} > \xi$. Hence, the alternative contracts make the firm ex ante better off, giving the desired contradiction. **Q.E.D.**

Proof of Proposition 6. The proof follows from the discussion in the main text. **Q.E.D.**