Economic Brief

November 2022, No. 22-45

When Should Employees Be Suspended Instead of Fired?

By Borys Grochulski

The economic theory of incentives explains why a worker who consistently underperforms must be fired. To respond to incentives, the worker must maintain a stake in the relationship with the employer. When the worker's stake runs out, the relationship must terminate. This article reviews recent research showing that this explanation is oversimplified. A temporary suspension of the worker is usually sufficient to rebuild the worker's stake, which allows the productive relationship to resume without terminating. The costs and benefits of suspending the worker, however, can be highly sensitive to the worker's and the employer's outside options. For this reason, similar jobs can have vastly different average job durations.

For many employers, when an employee's job performance declines, it's time to let that employee go. But while this may be the best course of action in many cases, the use of suspension might work out much better for both the employer and the worker. In this article, I discuss my recent work on when suspension might be the preferred path, including what types of jobs are best suited for suspension rather than termination.

Moral Hazard

The economic theory of incentives is concerned with information frictions, where an employer/principal lacks sufficient information to perfectly assess the job performance of their worker/agent. One such friction is known as moral hazard. It arises when the employer cannot fully monitor the worker's actions performed on the job. Suppose the employer observes the worker's performance only as a sum of the worker's true effort and some external, random noise. Observed performance can be strong if the random noise shock happens to be favorable, even if the worker's effort is low. Similarly, the worker might be putting in strong effort, but the employer can still see dismal results just because the random noise is unfavorable. Moral hazard arises as a temptation for the worker to shirk and blame random shocks for the poor observed performance.
To solve this moral hazard problem, the employer must ensure that shirking — although feasible for the worker at all times — is never in the worker’s own best interest. The worker's overall stake in the relationship is a key tool for meeting this requirement, known as the incentive compatibility condition. The worker’s stake is measured by the excess of the value she is owed by the employer over her outside value. Incentive compatibility is achieved by exposing the worker's stake to performance risk.

**Exposure to Own Performance Risk**

Suppose shirking gives the worker a private benefit that she values at $100, but it also lowers the performance observed by the employer (say, weekly sales) by $2,000. To discourage shirking, the worker's contract must ensure at least a 5 percent exposure of the worker's value in the relationship to the observed measure of performance (in this example, weekly sales).

With this exposure, shirking does not benefit the worker: By shirking, she can gain $100 as a private benefit, but since shirking also lowers the observed performance by $2,000, the value owed to her in the relationship goes down by $100 (5 percent of $2,000). It's a wash, so the worker might as well choose not to shirk. With this exposure, the overall contract is incentive compatible.

Maintaining an exposure of the worker's stake to performance risk implies that this stake must move up and down following random noise shocks. In our example, the 5 percent exposure to observed performance ensures the worker will choose not to shirk. But then, even without shirking, her stake in the relationship must absorb 5 percent of every realization of random noise shocks, as these shocks impact the observed performance one-for-one.

Ideally, the worker would be insured against the impact of random noise shocks on her performance. But with the employer being unable to tell noise from shirking, maintaining a 5 percent exposure to shirking produces as a necessary byproduct the same exposure of the worker's stake to random noise.

With this exposure, the worker's stake (or her "skin in the game") can have quite complicated dynamics. Her stake increases if observed performance is better than expected and decreases if the performance is below par. In particular, a long enough streak of negative random noise shocks can — despite the worker's best efforts — reduce the worker's stake in the relationship down to nil.

**A Poverty Trap**

When the worker's skin in the game runs out — that is, her value in the present relationship becomes equal to her value of finding another job — the employer faces a difficult problem: Incentives can no longer be provided to the worker, because she no longer has anything to lose in the present relationship.
Indeed, continued exposure of the worker's stake to performance risk requires that her stake be lowered further if performance continues to be subpar. But doing so would push the worker's value strictly below her value of finding another job. In this case, the worker would simply leave the current job and take her outside option instead, which effectively disarms the employer's incentive scheme. In many situations, this problem is impossible to overcome. That is, the relationship ends (meaning the worker is fired) as soon as the worker's stake runs out.\textsuperscript{2}

At face value, this outcome seems paradoxical: In pure productivity terms, the worker remains as good as she was at the start of the relationship, and, having worked under an incentive compatible contract, she never shirked. Yet, to ensure an overall incentive compatibility of the long-term relationship, the employer must at this point stick to the letter of the contract and end the relationship. This outcome is akin to a poverty trap: An otherwise productive worker can no longer be trusted to exert effort because she does not have any value she can expose to the risk of her own future performance.

**Escaping the Trap via Suspension**

Recent research, however, has discovered additional contractual possibilities that can overcome the poverty trap problem in many circumstances. In a forthcoming paper "Termination as an Incentive Device" — co-authored with Yuzhe Zhang and building on the 2013 paper "Optimal Contracts with Shirking" by John Zhu — we study temporary suspension of the worker as an alternative to terminating the relationship upon the worker's skin in the game running out.

Suspension is a contractual phase where the worker is asked to provide no effort (that is, the contract actually asks the worker to shirk) while receiving no compensation from the employer. The worker's exposure to performance risk is eliminated in suspension, which makes the shirking incentive compatible.

**Underpayment to Build a Stake**

In suspension, the flow of value the worker receives is restricted to her private benefit from shirking. How this value relates to the flow value of the agent's outside option becomes key for whether suspension can rebuild the worker's skin in the game. If her value of shirking is low, suspension temporarily underpays her, which means the value owed to her by the employer goes up.

To see this in a concrete example, suppose the worker's value from shirking matches exactly the flow value of her outside option. With her skin in the game at zero, her outside option is the same as the value owed to her by the employer. Therefore, the value of shirking matches exactly the flow equivalent of the value owed to the worker under the contract, keeping this value unchanged during suspension.\textsuperscript{3} In this example, suspension
does not rebuild the worker's stake in the relationship. Suspending the worker, thus, would do no good. It would only keep the worker stuck, providing no effort and remaining with zero stake in the relationship forever.

Suppose, however, that the worker's value from shirking is below the flow value of her outside option. Since the outside option is at this point the same as the value owed to the agent in the relationship, the value of shirking is lower than the flow equivalent of the value owed to her under the contract. Suspension now does rebuild the worker's stake. Indeed, the private benefit of shirking does not keep up with what the employer owes the worker. Thus, the employer effectively underpays the worker during suspension, and, hence, the worker must be promised a higher value in the future (that is, her stake increases). In this case, thus, suspension is a viable alternative to termination when the worker has no skin left in the game.

It may be convenient here to think of the total value owed to the worker in the relationship as the balance of a bank account, denoted by $L$, and the value from shirking during suspension as a stream of distributions from the account paid out to the agent over time. Suppose the bank account earns a constant rate of return of $r$. If the stream of distributions equals exactly $rL$, the balance of the account remains constant (at $L$) forever, meaning the worker's stake does not grow. But if the distributions are less than $rL$, the undisbursed return is capitalized into the account's principal balance, and the worker's stake grows.

As we see, suspension is effective as a tool for rebuilding the worker's stake in the relationship if her private benefit from shirking is low relative to her outside option. However, even when suspension can do the job, we also must ask if it is cost effective enough to be the better option than terminating the relationship after poor performance.

**Costs and Benefits to the Employer**

To examine the relative value to the employer of suspending the agent rather than terminating, we need to understand the employer's costs and benefits of the two options. In our paper, we describe this trade-off and study how the outside options of both employer and the worker affect it.

Suspending the worker entails an upfront cost to the employer and a delayed benefit. The cost comes from the fact that the worker is asked for no effort during her suspension, making the relationship temporarily unproductive. Clearly, the more damaging the worker's shirking is to the employer's bottom line, the lower the value of suspension to the employer.

The benefit of suspending rather than terminating is the productivity of the relationship regained as soon as suspension ends. The magnitude of this benefit depends on two factors:

- How quickly the suspension phase can end
• How productive the relationship is outside of suspension

Outside Options

The impact of the employer's outside option on this trade-off is straightforward: The employer's outside option always increases the value of terminating, making suspension relatively less attractive. Indeed, if the employer's outside option improves, the cost of having the worker shirk for the duration of suspension becomes relatively higher, which makes the cost of suspension relatively higher.

The impact of the worker's outside option on the employer's value of suspension is twofold. If the worker's outside option is high, on the one hand, the private benefit of shirking she enjoys during suspension is relatively low, which allows suspension to rebuild her skin in the game quickly. A lower expected duration of suspension — that is, the amount of time the worker must be suspended for — makes suspension a more attractive tool for the employer. On the other hand, a high outside option for the worker makes the relationship outside of suspension overall less profitable for the employer, as it decreases the room for the provision of incentives in normal times, when the worker is not suspended.

Testable Implications

These two opposing forces have interesting testable implications: Suspension should be most useful in mid-level jobs, where the workers' outside options are neither too high nor too low. CEOs, for example, would not be good candidates for suspension, as their outside options are high. Workers in occupations with low productivity and low compensation should not be good candidates for suspension either because their outside options are not much higher than the value of remaining on the job while accepting prolonged or repeated episodes of suspension.

Workers in mid-level jobs, however, could be good candidates for suspending rather than terminating, as their outside options are moderate — which means suspension can be relatively swift — and the value to the employer of a restored productive relationship is significant. With this use of suspension, thus, average job durations of workers in mid-level roles can be longer than those of both bottom-level and top-level roles.

The decision to suspend or terminate the relationship is a discrete one. Given two occupations of similar productivity, it may be optimal to use suspension with one but termination with the other simply because the outside options (to either the employer or the agent) are different. In such cases, the observed job durations may be very different despite the two jobs being otherwise very similar.

In our model, suspension has a narrow meaning of "do nothing and get paid nothing." But in practice, suspension can take many forms, as long the worker is underpaid and the incentives are switched off. That is, suspension can mean a reassignment away from tasks loaded with moral hazard toward tasks free of moral hazard or with a low moral hazard
component. In academia, for example, a professor may be assigned a high teaching load or a lot of administrative work following a period of low output in research. After such a spell of easily verifiable work, the riskier and harder-to-verify research or creative effort can resume.

**Further Research Questions**

Robustness of the optimality of suspension to changes in several aspects of the relationship presents some questions for further investigation. In a job with multiple tasks, which tasks would be best assigned during suspension? How does the use of suspension depend on the employer's impatience? Would a more patient employer be willing to rely on suspension more and termination less? Can some form of suspension be useful in providing incentives to teams of agents, where the moral hazard problem applies to the whole group?

Borys Grochulski is a senior economist in the Research Department of the Federal Reserve Bank of Richmond.

---

1. *These dynamics are particularly interesting in models where the worker can choose different levels of effort intensity.*


3. *The flow equivalent is defined as a permanent payment equal in value to a given lump sum. Specifically, for a worker discounting the future at the rate of time preference $r$, the flow equivalent of a lump sum $L$ is $rL$.*

---

*To cite this Economic Brief, please use the following format:* Grochulski, Borys. (November 2022) "When Should Employees Be Suspended Instead of Fired?" Federal Reserve Bank of Richmond Economic Brief, No. 22-45.

This article may be photocopied or reprinted in its entirety. Please credit the author, source, and the Federal Reserve Bank of Richmond and include the italicized statement below.

*Views expressed in this article are those of the author and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.*

---
Subscribe to Economic Brief

Receive an email notification when Economic Brief is posted online:

Email address

Subscribe

Contact Us

RC Balaban
(804) 697-8144

© 1997-2022 Federal Reserve Bank of Richmond