# Multitasking Incentives and Biases in Subjective Performance Evaluation*

Shingo Takahashi†
Hideo Owan‡
Tsuyoshi Tsuru§
Katsuhito Uehara¶

Subjective performance evaluation serves as a double-edged sword. While it can mitigate multitasking agency problems, it also opens the door to evaluators' biases, resulting in lower job satisfaction and a higher rate of worker quits. Using the personnel records of individual sales representatives in a major car sales company in Japan, we provide novel evidence for both sides of subjective performance evaluation: (1) the sensitivity of evaluations to sales performance declines with the marginal productivity of hard-to-measure tasks, and (2) a within-worker decline in evaluation that is unrelated to observable performance is consistently associated with higher worker quits despite our attempts to mitigate endogeneity bias.

*JEL Classification*:M52, M55

# 1. Introduction

An incentive contract that ties compensation to observable performance measures can provide strong incentives (Paarsch and Shearer 1999, 2000; Lazear 2000; Haley 2003; Bandiera, Barankay, and Rasul 2005, 2007). Prior works have provided strong evidence of the link between pay and performance. However, these studies were limited to occupations and workplaces where the tasks are clearly defined and worker productivity is easily measured in all key dimensions including quality. Researchers in the social sciences have long recognized that a contract that is solely based on observable performance measures often produces dysfunctional responses.[1] For example, if car sales representatives are incentivized

*This research was supported by Grant-in-Aid for Scientific Research (A) No.2524504, (B) No.21330058, and (C) No. 22530216 from the Japan Society for the Promotion of Science (JSPS).

†Corresponding author, Graduate School of International Relations, International University of Japan, 777 Kokusai-cho, Minamiuonuma, Niigata 949-7277, Japan, Email: staka@iuj.ac.jp, Tel: 81-25-779-1507.

‡Institute of Social Science, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033, Japan, Email: owan@iss.u-tokyo.ac.jp, Tel: 81-3-5841-4985, Fax: 81-3-5841-4905.

§Institute of Economic Research, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8603, Japan, Email: tsuru@ier.hit-u.ac.jp, Fax/Tel: 81-42-580-8384.

¶Faculty of Human Studies, Tenri University, 1050 Somanouchi-cho, Tenri, Nara 632-8510, Japan, Email: uehara@sta.tenri-u.ac.jp, Tel: 81-743-63-7094, Fax: 81-743-62-1965.

[1]Some early works that compiled anecdotes of dysfunctional responses include Laurence and Hull (1969) and Kerr (1975).

only by commissions on profits, they may ignore other tasks such as mentoring junior sales representatives.

The nature of such problems was formally analyzed by Holmstrom and Milgrom (1991), who showed that when an agent performs multiple tasks, some of which are hard to measure, the incentivized agent allocates more effort to the measured tasks, and may neglect the unmeasured tasks. Following their work, these incentive problems are known as multitasking agency problems.[2] Another related problem is workers gaming the system when presented with short-term incentives by manipulating the performance measure itself (Healy 1985; Holthausen, Larcker, and Sloan 1995; Oyer 1998; Owan and Tsuru 2011; Larkin 2014).

Multitasking agency problems arise because some tasks are hard to measure. Thus, one way to mitigate these problems is to use subjective performance evaluations (Prendergast 1999). For example, even if the mentoring of junior sales representatives is hard to measure, supervisors may still be able to subjectively assess such tasks. A sizable theoretical literature has analyzed the use of subjective evaluation in an incentive contract. By their nature, subjective measures attempt to capture aspects of performance that are not verifiable by a third party. Thus, the earlier literature focused on the conditions under which non-verifiable measures can be incorporated in an incentive contract (Bull 1987; MacLeod and Malcomson 1989; Pearce and Stacchetti 1998). Since subjective evaluation includes inherently private information, more recent literature analyzed the consequences of incorporating performance measures that are based on private opinions in an incentive contract (Baker, Gibbons and Murphy 1994; Levin 2003; MacLeod 2003; Fuchs 2007; Chan and Zheng 2011).

In all theoretical studies, the underlying premise is that subjective evaluation is used to incorporate hard-to-measure tasks in an incentive contract. By hard-to-measure tasks, we mean tasks that are hard to quantify in a way that can be rewarded with formulaic

[2] Baker (2002) also analyzed a similar problem in which the effects of the agent's action on the performance measure differ from its effects on the firm value.

bonuses. A natural empirical question is whether subjectivity is indeed used for this purpose. However, only a few studies have examined this question. By using branch managers' compensation data from 150 auto dealerships, Gibbs et al. (2004) showed that dealerships that face greater multitasking agency problems or a greater threat of gaming behaviors are more likely to use discretionary bonuses. Using CEO compensation data, Bushman, Indjejikian, and Smith (1996) showed that growth opportunities and product development cycles, which are proxies for multitasking agency problems, are positively related to the use of individual performance evaluations.[3] Hayes and Schaefer (2000) showed that performance measures that are unobserved by third parties are indeed used in setting CEOs salaries.[4]

Although subjective evaluation is useful in providing multitasking incentives, it inevitably opens the door to evaluators' biases, such as favoritism and discrimination, that cause inefficiencies. Prendergast and Topel (1993) showed that such biases lower the utility of risk-averse workers and distort promotion decisions. MacLeod (2003) developed a model that allowed a supervisor's assessment of a worker's performance to deviate from the worker's own assessment and derived an optimal contract that depends on the reported assessment from both sides. MacLeod termed this deviation the "perceived bias" and argued that perceived bias would cause conflicts that reduce the future output.[5] Similarly, Levin (2003) demonstrated that the use of subjective measures inevitably causes conflicts due to differences in opinions which may motivate workers to quit. A distortion in subjective per-

[3]Murphy and Oyer (2003), however, did not find this evidence.

[4]Some studies provide evidence that the link between pay and easily measured tasks becomes weak or less explicit when there are some concerns about multitasking agency or gaming problems, although they do not show direct evidence of the use of subjective measure on hard-to-measure tasks. For example, Hoppe and Moers (2011) found that firms with greater environmental uncertainty tend to write CEO contracts free of formulaic bonuses. Ederhof (2010) showed that subjectivity is more likely to be used when contractible outcomes are either high or low.

[5]Prendergast and Topel (1993), Levin (2003), and MacLeod (2003) also predicted another type of bias in subjective evaluations, supervisors not sufficiently distinguishing between workers, known as centrality bias. Empirical evidence of this type of bias can be found in Murphy and Cleveland (1991) and Larkey and Caulkins (1992).

formance evaluation may also come from collusion between a manager and her subordinate, and preventing such collusion is costly (Thiele 2013).

Subjective performance evaluation is therefore a double-edged sword. It can mitigate multitasking agency problems, but it opens the door for bias, resulting in greater conflicts. Thus, the purpose of this study is to provide evidence for both sides of subjective performance evaluation.

Our first goal is to provide new and straightforward evidence that subjectivity is indeed used to incorporate hard-to-measure tasks in an incentive contract. We use personnel and transaction records of new car sales representatives in a major car sales company in Japan, which we have given the pseudonym "Auto Japan." In Auto Japan, new car sales representatives work under commission, which alone provides a strong incentive to perform well. However, Auto Japan also conducts annual performance reviews in which the supervisors subjectively rate workers' performance on a 5-grade scale. The evaluation results are then used to determine annual salary raises. According to our interviews with several branch managers, the reason they conduct the performance reviews is to reward effort on tasks not captured by the sales figures. After several interviews, we identified two important hard-to-measure tasks: (1) mentoring junior workers and (2) building long-term customer relationships. We provide evidence that subjectivity is indeed used to reward these hard-to-measure tasks.[6]

Our empirical strategy is as follows: consider a supervisor in a car dealership who needs

[6]It is worth mentioning that this analysis relates to a growing literature that investigates how workers change effort allocation in response to changes in compensation plans. Drago and Garvey (1998) showed that when promotion incentives are strong, individual effort increases while helping effort decreases. Using physicians' compensation data in Canada, Dumont et al. (2008) showed that the move away from a fee-for-service compensation plan, where remuneration is tied to the quantity of care, towards a flatter compensation plan increased the time physicians spent on each patient, as well as the time they allocated to teaching and administrative work. Using a personnel data set from a multinational law firm, Bartel, Cardi, and Shaw (2012) showed that the firm's move from individual incentives towards leadership incentives reduced team leaders' billable hours and increased their non-billable hours. While these studies focused on how workers' effort allocation might change in response to a change in weights among competing tasks, our study focuses on how supervisors might alter these weights given the presence of multitasking problems.

to motivate sales representatives to not only sell more, but also to do hard-to-measure tasks such as mentoring junior workers. If the supervisor incorporates sales representatives' mentoring performance in his or her evaluations, the weight placed on sales would decline as the weight on mentoring increases, leading to a decrease in evaluations' sensitivity to sales performance. We use this prediction for our test of multitasking incentive provisions.

The second goal of this study is to examine the effects of evaluation bias on worker quits–one outcome of conflicts between supervisors and workers. Some studies have pointed to the existence of biases in subjective evaluation (Goldin and Rouse 2000; Elvira and Town 2001). However, the effects of biases on worker quits are largely untested.[7] The key issue is how to identify evaluation bias. To obtain evidence suggesting the effect of evaluation bias on worker quits, we constructed two measures of variations in evaluations that are unrelated to observed performance measures: unexplained evaluation gaps computed based on residuals, and within-worker change in the supervisor-worker match fixed effects.

The first measure is based on the idea that if evaluation bias exists, it should appear in the residuals of the evaluation regressions. Thus, we constructed an indicator of a negative evaluation gap, which takes the value 1 if the negative residual is large enough to lower the evaluation grade by one level (e.g., a B grade when an A is expected). One problem with this approach is that this measure reflects the following three factors: (1) taste-based evaluation bias, (2) perceived bias, and (3) workers' performance on hard-to-measure tasks that are unobservable to researchers. Obviously, the presence of the third factor causes a difficulty in isolating the effects of evaluation biases (the first and second factors) on worker quits.

If the third factor is time-invariant, then we can eliminate the effect of the third factor

[7]Giuliano, Levine, and Leonard (2005) found that having supervisors of races different from that of their subordinates increases the probability of quits for black and Hispanic workers. This result suggests the existence of bias, but it is not conclusive. Engellandt and Riphahn (2011) evaluated the effects of favoritism-tainted evaluations on worker productivity, but they did not test the effects of biased evaluations on worker quits.

by simply including worker level fixed effects in a quit regression. However, if there is a time-varying component in the third factor, which almost surely is the case, this approach still faces an endogeneity bias problem. We attempt to reduce this endogeneity bias by using the change in evaluation that is caused by the discontinuous shift in the supervisor-worker relationship. This is measured by the within-worker change in the supervisor-worker match fixed effects.

The basic assumption of the latter approach is that evaluation bias is relatively time invariant within supervisor-worker match, but it varies significantly across different supervisor-worker matches. Thus, a change in supervisor would cause a large change in evaluation bias. However, the third factor would not shift much upon a change in supervisor unless the worker was re-assigned to different tasks at the same time, but this is unlikely in the sales department because their primary task is always selling cars. Under this assumption, the within-worker change in supervisor-worker match fixed effects should largely pick up the effect of evaluation bias.

Although we admit that this might not completely eliminate endogeneity biases, it should at least mitigate such biases, allowing us to provide suggestive evidence of the inefficiencies caused by evaluation biases. More specifically, if the latter approach using the change in supervisor-worker match fixed effects produces a similar result to the first approach using the unobserved evaluation gap, this should mean that the effect of an unexplained decline in evaluation comes largely from evaluation bias.

To preview the results, both approaches produced very similar estimates: an unexplained decline in evaluation increases quit probability by roughly 5 percentage points. The estimated impact that is robust to the change in approach seems to suggest that taste-based bias or perceived bias play a substantial role in inducing worker quits.

The validity of our inference is further enhanced by additional analysis using the information from an employee satisfaction survey which include a question of how fair they believed evaluation results were. We showed that an unexplained negative evaluation gap is significantly associated with the worker's feeling of unfairness in the evaluation results.

The rest of the paper is structured as follows: Section 2 presents two standard models used to derive our empirical prediction. Section 3 outlines Auto Japan's performance evaluation systems and possible sources of multitasking agency problems. Section 4 describes the data and Section 5 discusses the results and their implications. Finally, Section 6 concludes with a summary of our results and remaining issues.

## 2. Theoretical Background

We motivate our empirical strategy using the standard results from a simple multitasking principal-agent model (Baker, 2002), where the agent performs two tasks. If the marginal productivity of one task increases, the principal should induce the agent to reallocate effort towards that task. As long as these two tasks are substitutes in the agent's effort cost function, the principal should increase the weight for that task, and reduce the weight for the other task. We use these implications to develop a testable prediction.

For simplicity, we assume that the performance of both tasks is contractible. However, contract theory literature generally finds that when a performance measure is observable to both the principal and the agent (or their beliefs are perfectly correlated), they can achieve the same outcome that would be obtained under a verifiable performance measure (Bull 1987; MacLeod and Malcomson 1989; Pearce and Stacchetti 1998; MacLeod 2003). In short, the verifiability assumption is not actually required. Loosening the assumption is critical here because performance is unlikely to be verifiable on either hard-to-measure task we discuss in our empirical section.

Let $t_s$ and $t_m$ be the effort spent on sales and mentoring tasks, respectively. Let $W$

be the wage for the agent. The value of the firm is then given by: $V$=$B_1t_s$+$B_2t_m$+$\varepsilon$-$W$ where $B_1$ is the marginal productivity of sales effort and $B_2$ is the marginal productivity of mentoring effort. The term $\varepsilon$ captures the environmental uncertainty the firm faces with mean zero.

The supervisor evaluates both tasks, which are given as: $p_s = t_s + \eta_s$ and $p_m = t_m + \eta_m$. The terms $p_s$ and $p_m$ are the measures of sales and mentoring effort, respectively. The terms $\eta_s$ and $\eta_m$ are the random errors in measuring performance; these are assumed to be normally distributed with variances $\sigma_s$ and $\sigma_m$. For simplicity, we assume that these errors are independent.

Consider the linear compensation scheme $W = \beta_0 + \beta_1 p_s + \beta_2 p_m$. Assuming that the agent's cost of effort function is C($t_s$,$t_m$)=$at_s^2 + bt_m^2 + ct_st_m$, and that the agent has the exponential utility function U=-exp(-r(W-C)), the optimal weights are given by:

$$\begin{aligned}
\beta_1^* &= \frac{(1+rb\sigma_m^2)B_1 - rc\sigma_m^2 B_2}{1+rb\sigma_m^2+ra\sigma_s^2+r^2(ab-c)\sigma_s^2\sigma_m^2} \\
\beta_2^* &= \frac{(1+ra\sigma_s^2)B_2 - rc\sigma_s^2 B_1}{1+rb\sigma_m^2+ra\sigma_s^2+r^2(ab-c)\sigma_s^2\sigma_m^2}
\end{aligned}$$

It is natural to assume that sales effort is more tiring if one is also mentoring junior workers. Therefore, we assume that sales and mentoring are substitutes in the agent's cost function so that $c > 0$. The denominators in the formula above are positive due to the second order condition. Thus, the above equations have the following two implications:

**Implication 1**: $\beta_1^*$ (the weight for the sales effort) is a decreasing function of $B_2$ (the marginal productivity of the mentoring effort).

**Implication 2**: $\beta_2^*$ (the weight for the mentoring effort) is an increasing function of $B_2$ (the marginal productivity of the mentoring effort)

Regarding Implication 1, Holmstrom and Milgrom (1991) demonstrated that even if the principal does not measure the hard-to-measure task at all, the weight for the sales task is still a decreasing function of $B_2$ in order to reduce effort misallocation.[8] Showing that $\beta_1^*$ is a decreasing function of $B_2$, thus, does not necessarily indicate that subjective evaluation is actually used to measure the performance in hard-to-measure tasks. Such a result, however, is clear evidence that the firm is adjusting pay to deal with multitasking agency problems.

Because we cannot directly measure $\beta_1^*$ and $\beta_2^*$, we need to press on a little further to obtain a testable prediction. Note the following derivations:

$$\frac{\partial W}{\partial B_2}|_{p_s=\widetilde{p}} \quad = \quad \frac{\partial \beta_0^*}{\partial B_2} + \frac{\partial \beta_1^*}{\partial B_2}\widetilde{p} + \frac{\partial \beta_2^*}{\partial B_2}p_m^* + \beta_2^* \frac{\partial p_m^*}{\partial B_2}$$

Now, by plugging in the optimal mentoring effort, $t_m^*$, in $p_m^*$ as $p_m^* = t_m^* + \eta_m$, we obtain:

$$\frac{\partial W}{\partial B_2}|_{p_s=\widetilde{p}} = (\frac{\partial \beta_1^*}{\partial B_2} - \frac{c}{2b}\frac{\partial \beta_2^*}{\partial B_2})\widetilde{p} + \frac{\partial \beta_0^*}{\partial B_2} + \frac{\beta_2^*}{b}\frac{\partial \beta_2^*}{\partial B_2} + \widetilde{\varepsilon}$$

where $\widetilde{\varepsilon} = \frac{\partial \beta_2^*}{\partial B_2}(\frac{c}{2b}\eta_s + \eta_m)$. Note that $\frac{\partial^2 W}{\partial B_2 \partial p_s} = \frac{\partial \beta_1^*}{\partial B_2} - \frac{c}{2b}\frac{\partial \beta_2^*}{\partial B_2} < 0$. Therefore, pay sensitivity to sales performance declines with $B_2$ because of two reasons. First, the optimal weight on sales performance declines. Second, the optimal weight for mentoring effort increases. We can therefore make the following prediction.

**Testable Prediction**: The sensitivity of wage to sales performance declines with the marginal productivity of other hard-to-measure activities. Namely, $\frac{\partial^2 W}{\partial B_2 \partial p_s} < 0$.[9]

[8] This is the case when $\sigma_m$ goes to infinity in the above model (in such a case, $\beta_2^*$ will be set equal to zero).

[9] We can extend this prediction to the case in which the supervisor's assessment of the worker's performance is not perfectly correlated with the worker's self-evaluation. MacLeod (2003)'s theory implies that as the assessments become less correlated, the partial pooling expands in the equilibrium, which makes the pay-performance linkage weaker. Note that the multitasking problem could be a primary source of imperfections in the correlation between the supervisor's assessment and the worker's self-assessment. On the one hand, when the worker has a single task and the performance of the task is precisely measured, both assessments are likely to be perfectly correlated. On the other hand, when the worker has other duties that are hard to measure, the supervisor's and the worker's assessments are likely to diverge. Thus, when the multitasking problem is severe (i.e., $B_2$ is large), the correlation between the supervisor's assessment and the worker's self-assessment is generally weak, and therefore, the supervisors are less likely to distinguish between superb performance and mediocre one, leading to a weakening of pay-to-sales-performance sensitivity. This is the same prediction as derived from the standard multitasking agency model.

# 3. Performance Evaluation System, Practices, and Empirical Strategy

## 3.1 Auto Japan's Performance Evaluation System

Auto Japan uses both objective and subjective performance evaluations in determining compensation for each salesperson. First, gross profit earned determines the commission component, $C$, the worker receives. Subjective performance evaluation affects the base salary, $W$, as explained below. The worker's total monthly pay, $P$, is the sum of $C$ and $W$. Although there was a change in how the commission is calculated in 2001, there was no fundamental change in how base salary was updated during our observation period. The change in the sensitivity of pay-for-performance in 2001 should not affect our results because year effects are fully controlled for in all of our analyses.[10]

Sales representatives at Auto Japan are classified into one of five "salary stages"–S, A, B, C, and D. A salary stage is determined by each worker's base salary level depending on which pay range it belongs to. For example, if the base monthly salary is above 200 thousand yen, the sales representative was in salary stage S for the period of our observation.[11]

Base salary annual raises are determined by an annual performance evaluation conducted at the end of each fiscal year, and each worker is given an evaluation grade of s, a, b, c, or d. At the beginning of the year, a sales representative and his supervisor hold a discussion to set goals in both quantity and quality dimensions of their work.[12] The quantity dimension includes (1) number of cars sold, (2) annual profit earned and (3) profit earned from insurance sales. The quality dimension includes (4) number of days required to collect payment, and (5) percentage of former customers who brought their cars in for

[10] More detailed information of Auto Japans compensation policy and the pay policy change in 2001 can be found in Owan and Tsuru (2011) or provided upon request.

[11] Stage D indicates a base salary between 40 to 80 thousand yen; Stage C is for a base salary between 80 to 120 thousand yen; Stage B, 120 to 160 thousand yen; and Stage A, 160 to 200 thousand yen.

[12] We use the masculine pronouns "his" and "him" throughout the paper because there were only several female managers and representatives in our samples.

inspection. Note that (4) and (5) are included to provide incentives to attract customers who are financially sound and to enhance customer satisfaction.

Branch managers also have the discretion to add goals in other dimensions. For example, if a branch manager wants to encourage salespersons to develop new corporate customers, he can set a goal for the number of new corporate accounts each salesperson should open. At the end of the year, the supervisor interviews each sales representative individually to discuss the degree to which his goals were achieved. The annual performance evaluation is determined based on these interview results.

Yearly changes in base salary are determined by the "wage raise matrix" shown in Table 1, in accordance with evaluation grades and salary stages. Formally, we can express this relationship using the equation $\Delta W_{t+1} = f(Evaluation_t, Salary\_stage_t) = \tilde{f}(Evaluation_t, W_t)$ where the salary stage variable is replaced by the current wage because the former is automatically determined by the latter. As shown in Table 1, a worker needs to obtain an evaluation equivalent to his salary stage (i.e. s to S, a to A, etc.) or above to receive a pay raise. An evaluation that is lower than one's salary stage leads to a wage cut. In other words, any deviation of the evaluation from a worker's salary stage has a strong implication for wages the following year.

Importantly, according to Auto Japan's stated evaluation procedures, all sales representatives are held to the same standards. That is, workers with equal performance are supposed to receive the same evaluation regardless of salary stage. Despite this stated policy, supervisors may use salary stages as "reference grades", and decide how much to deviate from them if he incurs some psychic cost when giving a subordinate a grade that leads to a wage cut. Therefore, the actual evaluation may be systematically biased.

## 3.2 Multitasking Agency Problem

This section outlines possible sources of multitasking agency problem in Auto Japan. The main purpose is to identify proxies for the marginal productivity of hard-to-measure tasks. Auto Japan is an auto dealership and, unlike other white-collar jobs, sales representatives have readily available, objective performance measures–the gross profit earned by each representative.[13] Why, then, does Auto Japan not incentivize sales representatives solely with commissions? Figure 1 illustrates other tasks that might need to be taken into account in providing incentives. First, the management may care more about the quality of sales activities than salespeople do because of obvious externalities–for example, good customer care by individual sales representatives helps build the reputation of the company.

The other two tasks illustrated in Figure 1 are what we learned from interviews with several branch managers and one executive. They said that subjective performance evaluation is needed because sales figures miss some important tasks performed by the sales representatives: mentoring junior sales representatives and building and sustaining long-term customer relationships.

Mentoring junior sales representatives is the most frequently mentioned reason for the use of annual performance evaluations. If sales representatives are incentivized solely based on the profits they earn, mentoring tasks will be neglected. The annual performance evaluation is used to reward mentoring. The following excerpt from our interview with a branch manager effectively illustrates this:

"Suppose that there are sales representatives A and B. Representative A sells 10 cars per month, and at the same time takes good care of junior representatives, sitting beside them and helping them negotiate deals and giving advice in the various phases of sales

[13]Profit is a better measure of performance than the number of vehicles sold because the profit margin differs across car models, and sales representatives are encouraged to increase profits by selling car accessories and insurance.

activities. Representative B also sells 10 cars per month but he does things only for himself. Representative A will definitely receive a much higher evaluation at the performance review. Some of these junior representatives will eventually become supervisors themselves and educating them is absolutely necessary."[14]

As a proxy for the marginal productivity of mentoring efforts, we use the ratio of junior representatives to experienced representatives in a sales group because we do not have information about exactly who provided mentoring. Seventy-four Auto Japan branches have new car sales departments, each of which typically has one or two sales groups with its own supervisor. We define a junior sales representative as a representative in his first year at Auto Japan, since a new entrant is likely to receive the most extensive mentoring. We define an experienced representative as a representative who is in the third year of tenure or greater.[15] This definition is derived from our interview with an executive who stated that it takes approximately three years for a worker to be able to perform all sales tasks competently. Thus, we construct this variable as:

$$\begin{aligned} &\textit{Junior to experienced rep ratio}_{it} \\ &= \begin{cases} \dfrac{\#\textit{junior sales reps. in the sales group}_{it}}{\#\textit{ of experienced sales reps. in the sales group}_{it}}, & \text{for experienced sales reps} \\ 0, \ \text{for those \underline{not} classified as `experienced reps'} \end{cases} \end{aligned} \tag{1}$$

We interpret this variable as the average number of junior sales representatives that each experienced representative has to mentor in the sales group, or the probability of being assigned to a junior representative as a mentor. We set this variable to be zero for those who are not classified as 'experienced'. As Table 2 shows, the average *Junior to experienced rep ratio* is 0.046. Note that our proxy variable for mentoring is constant for all experienced representatives within a sale group.

The need to build long-term customer relationships is another frequently cited reason

[14]According to our interview with a branch manager in May 2006.

[15]This definition assumes that the second year representatives neither receive nor provide mentoring.

why annual performance evaluations are needed in addition to commission payments. If sales representatives maintain good relationships with their former customers, these customers are more likely to bring their cars to Auto Japan's maintenance department and more likely to purchase another car from Auto Japan in the future. For this reason, Auto Japan encourages sales representatives to periodically visit or call their customers who have purchased cars from Auto Japan. Because the maintenance departments' profits do not directly benefit the salespersons, the externality is not internalized. In addition, given that sales representatives may be too myopic to foster customer loyalty that may not pay off for several years, it makes sense to provide additional incentives through subjective evaluation.

In order to test the multitasking incentive provisions, we need *variation* in the importance of long-term customer relationships. We argue that the efforts to build long-term relationships have a higher marginal productivity for corporate customers than for individual customers since corporate customers buy multiple vehicles and thus replacement demand and car inspection demand arise more frequently. In addition, sales to corporate customers are more likely to be the results of teamwork and thus free-riding may be of some concern–another reason why additional incentives via subjective evaluation is important for corporate customer sales.

Therefore, we use the share of total gross profit generated by corporate customers for each sales group to measure the importance of building long-term customer relationships. Our transaction data contain basic information about each transaction including the name of salesperson who sold the car, the name of the customer, and gross profit earned from the sale. From the customer name, we can roughly distinguish corporate customers from individual ones. Thus, we construct the following variable:

$$
\begin{aligned}
&(Corporate\ customers\ share) \\
=\ &\frac{(Annual\ profit\ from\ corporate\ customers\ in\ each\ sales\ group)}{(Total\ annual\ profit\ earned\ in\ each\ sales\ group)} \qquad (2)
\end{aligned}
$$

We expect that the efforts to build long-term customer relationships are more important if the share of corporate sales in the total profit is greater. Thus, we use this variable as a proxy for the marginal productivity of the customer relationship management task.

### 3.3 Empirical Strategy

The theoretical implication we need to test is whether the sensitivity of wages to sales performance declines with these proxies. This empirical implication from a static model is hard to test using the actual wage data in our sample because the wage level itself reflects the accumulated effects of past performance as the wage increase is tied to the worker performance in the previous period in accordance with the company's wage raise matrix described earlier. Therefore, it is more natural to examine the sensitivity of evaluation (or equivalently the wage increase) to sales performance instead. Therefore, we estimate the following equation:

$$\begin{aligned} Evalation^*_{it} &= \alpha_1 (Profit)_{it} + \alpha_2 (Profit)_{it} (\textit{Junior to experienced rep ratio})_{it} \\ &+ \alpha_3 (\textit{Junior to experienced rep ratio})_{it} \\ &+ \alpha_4 (Profit)_{it} (\textit{Corporate customers share})_{it} \qquad (3) \\ &+ \alpha_5 (\textit{Corporate customers share})_{it} + X_{it}\beta + (Branch\ fixed\ effects) + u_{it} \end{aligned}$$

where $Evaluation^*_{it}$ is the latent variable for the actual evaluation that takes value from 1, 2, 3, 4, 5. This is a conversion from the evaluation letter grade where the highest grade of S corresponds to 5, and the lowest grade of D corresponds to 1. $Profit_{it}$ is the annual gross profit earned by the $i^{th}$ worker in the year $t$. $X_{it}$ is a vector of other determinants of an evaluation–characteristics of the worker and the supervisor. As will be detailed later, we include branch fixed effect to control for the potential endogeneity of our marginal productivity measures. The coefficients of interest are $\alpha_2$ and $\alpha_4$. If our prediction holds, both $\alpha_2$ and $\alpha_4$ should be negative.

# 4. Data, Variables, and Descriptive Statistics

## 4.1 Data Set

We use three data sets we obtained from Auto Japan. The first data set contains detailed information about car sales made during fiscal years 1999 to 2004[16], such as the gross profit obtained from each sale and customer names. The second data set contains employee information including the year salespeople were hired and which branches and sales group they worked at particular points in time during our sample period. In addition, we are able to identify the supervisor of each sales group in each branch at particular points in time. The third data set contains the annual evaluation results for each sales representative. We merged these three data sets using the worker identification numbers. As we obtained the performance evaluation data only for the period 2000 to 2003, our analysis is restricted to these four years.

Observations for sales representatives who left the firm before the end of a fiscal year were dropped from the sample for that year because these workers are not evaluated. We measure each worker's tenure at the end of each fiscal year. If workers entered the firm in the middle of a fiscal year, their initial year's observations were dropped if they had worked less than nine months at the time of evaluation because the supervisor was unlikely to have sufficient information to judge their productivity.[17] In addition, sales representatives over age 60 are also excluded from our analysis, as they were former employees rehired as fixed-term contract workers after reaching the mandatory retirement age of 60, and their performance is not evaluated.

Auto Japan also has a sales department specializing in fleet sales to large corporations and its own affiliated rental car company. The sales representatives in this department are

[16]Fiscal years at Auto Japan start in April and end in March. We obtained data from December 1998 to December 2005, so the complete fiscal years in this period are 1999 to 2004.

[17]This eliminated 33 observations.

excluded from our dataset because the sales activities, profit margins, and incentive schemes in the department are very different from those of other sales departments.[18]

The above sample selection criteria resulted in an unbalanced panel of 686 new car sales representatives in 74 different branches, in 120 sales groups, over the period 2000-2003, resulting in a total of 2148 representative-year observations.

## 4.2 Variables and Summary Statistics

Table 2 shows the descriptive statistics of the variables that we use for our estimation. The average of *Evaluation* is about 2.7 (between b and c). Ninety-three percent of the observations received either a, b, or c. The variable *Profit* is the annual gross profit each worker earned, in millions of yen, during the fiscal year–the average is 20.71 million yen (approximately US$200,000).[19] Most of the variables are self-explanatory, though some variables are motivated by existing theories of subjective evaluation and worker behavior and thus deserve some explanation.

Merchant (1989) and Gibbs et al. (2004) emphasized that subjectivity is used to filter out uncontrollable risks. For example, if a supervisor judges that the sales at his branch are low due to factors beyond the workers' control, such as the relocation of a branch to a temporary site, he may inflate their evaluations so as not to punish workers for their bad luck. Uncontrollable risks could be either systemic–affecting all branches in the same way–or idiosyncratic, limited to a single branch. Note that the managers can evaluate idiosyncratic risks more easily than systemic ones by comparing branch performance with the firm-wide average.

We attempt to capture idiosyncratic, uncontrollable risks at the branch level using the

[18] In many cases, the deals are negotiated directly between the clients and the car manufacturer. Sales to the affiliated rental car company, for example, do not generate any profits. For this reason, workers in this department do not receive commissions on profits from sales.

[19] There are 157,897 transactions recorded in our data for the new car sales representatives. The majority of these transactions are actual sales (90%). The rest of the transactions were lease contracts. Our variable, *Profit*, includes both sales and lease contracts.

variable *Branch-firm productivity differences*, defined as branch productivity minus the firm-wide productivity. Branch productivity is the profit per salesperson at each branch. Firm-wide productivity is defined similarly.[20] Gibbs et al. (2004) emphasized that supervisors may adjust evaluations only in response to negative shocks because filtering out positive shocks would cause the workers to cut back their sales efforts in good years (ratchet effect). To capture this asymmetry, we split the productivity differences into positive and negative parts in the estimation.

We control for the total number of sales representatives in a branch for the following reason. Within a branch, each sales representative has his given territory. When sale people leave their branches, their territories are reassigned within that branch. Therefore, a reduction in the total number of representatives typically leads to an increase in profit per representative. However, this rise in profit may not be perceived as better performance and the supervisor may reduce the weight on profit in evaluations. Inclusion of the total number of sales representatives at a branch alleviates this effect.[21]

We also control for two additional branch characteristics: the average worker tenure, and the standard deviation of worker tenure at each branch. Most employees at Auto Japan joined the firm right after finishing school and the hiring of mid-career workers is rare. Therefore average worker tenure represents the average sales experience in each branch, whereas the standard deviation of worker tenure captures the degree of age diversity in a branch.

There are two reasons why we include the average worker tenure. First, the average worker tenure captures the average human capital. More experienced workers know more about how to develop and maintain good relationships with corporate customers and how

[20]In the computation of profit per representative, first-year representatives are excluded so that branch turnovers do not affect the variable.

[21]Note that when a representative worked for a fraction of a fiscal year, the fraction is used for the computation of the number of representatives.

to coordinate with other departments to increase cross-selling. The benefits of these activities may not be captured in short-term sales figures but may be highly evaluated in the subjective evaluation. To the extent to which these skills are shared, junior salespeople will benefit from the presence of experienced representatives in their branches. Second, higher average worker tenure means smaller average age differences between sales representatives and their branch managers, which may imply more effective cooperation and coordination within a branch.[22] Subjective evaluations may improve if the sales representatives respond well to the branch managers' directions.

Inclusion of the standard deviation of worker tenure is the result of our consideration of social identity theory or self-categorization theory in organizational behavior.[23] If small age diversity enhances group identity, supervisors may attempt to reinforce that identity by giving similar grades to every member of the branch, especially by giving lenient grades to low performers.

To provide some preliminary ideas about how evaluation is related to a worker's performance, Figure 2 plots *Evaluation* against *Profit*. Although we observe a clear positive relationship, we can see that the same profit often translates into different evaluations as well.

# 5. Empirical Analyses

## 5.1 Do managers adjust their evaluation for multitasking agents?

Our main goal is to determine if a worker's evaluation takes into account the performance of hard-to-measure tasks such as mentoring and building long-term customer relationships. We test our main prediction by estimating equation (3). We first estimate ordered probit models, the results of which are shown in Table 3.

[22] See Zenger and Lawrence (1989) who found a negative effect of age diversity on communication.
[23] See Mannix and Neale (2005) for related discussion.

Model 1 is the most parsimonious baseline model. The interactive terms between *Profit* and *Junior to experienced rep ratio*, and *Profit* and *Corporate customer share* are both negative and significant at the 5 percent level. Thus, the results are consistent with our prediction, which states that the sensitivity of wages to sales performance declines with the marginal productivity of other hard-to-measure tasks. Model 2 adds worker, supervisor and branch characteristics. The coefficients for both interactive terms remain negative and significant. Although the coefficient for the first interactive term is only weakly significant, the results are generally consistent with our prediction. The weak significance of the result for *Junior to experienced rep ratio* may come from the lack of information about who are the actual mentors of junior representatives.

As noted earlier, Auto Japan's stated evaluation procedure calls for "equal evaluation for equal performance regardless of one's salary stage". Hence, we did not include salary stage dummies in the first two specifications. However, as we also noted earlier, the wage raise matrix in Table 1 indicates that supervisors may use the worker's salary stage as a reference grade, and then determine how much to deviate from it. If this is the case, a person's evaluation might be higher than his true performance simply because his salary stage is higher. To account for this possibility, we add salary stage dummies in Model 3. As expected, the coefficients for stage dummies monotonically increase from C to S.

Our main prediction is not fully supported in this model specification. Although the coefficient for the interactive term between *Profit* and *Corporate customer share* is still significant at the 1 percent level, the coefficient for the interactive term between *Profit* and *Junior to experienced rep ratio* became insignificant. This does not necessarily mean that the evaluators do not take into account the increased role of mentoring in giving evaluation grades. If the chance of being asked to mentor is correlated with worker productivity (which is reflected in the salary stage), the coefficients of salary stage dummies should be overestimated while that of the interactive term between *Profit* and *Junior to experienced rep ratio should* be underestimated.[24]

One concern we have is that the number of junior workers in a branch is likely to be influenced by unobserved branch characteristics. In Auto Japan, once sales representatives are assigned to a branch, they will move to another branch only when promoted to a supervisor position. Thus, a job vacancy occurs only when a branch expands, a representative quits, or a representative is promoted. The frequency of these events is likely to be influenced by unobserved branch characteristics. Similarly, the corporate customer share is likely to be influenced by the demographics of the neighborhoods surrounding the branch. To control for the potential endogeneity caused by correlations between these unobserved characteristics and our multitasking proxies, we include branch dummies in Model 4. The inclusion of branch dummies did not alter the results qualitatively, but improved the significance of the interactive term between *Junior to experienced rep ratio* and *Profit* to the 5 percent level, although the statistical significance of the other interactive term dropped to the 10 percent level (pval=5.3 percent).

Unobservable supervisor characteristics such as managing and training ability may also be correlated with sales group turnover and promotion rates, which affects *Junior to experienced rep ratio*. Thus, we include supervisor dummies in Model 5. The coefficient for the interaction between *Profit* and *Junior to experienced rep ratio* became insignificant (pval=12 percent). However, the coefficient for the interaction between *Profit* and *Corporate customer share* is still significant at the 10 percent level. Relatively low significance of

[24] Another concern is model misspecification. In this analysis, we assume that the coefficients of current performance are constant, but it is quite possible that, as time goes by, the supervisor accumulates more information about each worker's ability and potential, and thus lower weights will be given to measures of current performance. Because ability is more fully revealed for workers in the higher salary stages due to their longer tenure, salary stages may capture the decreasing weight on current performance over tenure. We can test this hypothesis by including the interaction between *Profit* and *Worker's tenure*. If the hypothesis is correct, the inclusion of the cross term will reduce the effect of salary stages. We have conducted this test in an unreported regression, and found that: (i) the coefficient for the additional cross term is small and insignificant, and (ii) all other coefficients (including salary stage dummies) are essentially unaffected. Therefore, this employee learning hypothesis cannot explain the salary stage effects.

the interactive terms in Model 5, however, is likely to be due to smaller within-manager variations of our two focal variables.

Now, let us examine some other results. Did supervisors adjust for shocks that affected all salespeople in the branch and were beyond their control? Across all the models, the coefficients for the negative components of *Branch-firm productivity differences* are negative and significant. The coefficient for the positive part is also negative for all the models, but is insignificant for some models. Also note that the coefficient for the negative component is much larger in absolute value than that for the positive component, indicating that it is mainly the negative shocks that are filtered out in evaluation. Interestingly, this result is consistent with Gibbs et al. (2004), who predicted that uncontrollable risks would be filtered out only when they are negative, although our results do suggest that supervisors may adjust evaluation downward when there are positive shocks.

One possible criticism of our results so far is that a negative coefficient of an interactive term in non-linear regression does not necessarily imply substitutability of the two variables, which is defined by the cross-derivative of the expected value of *Evaluation* (see Ai and Norton 2003). To answer to this criticism, we re-estimate the same models using tobit regression.[25] Tobit regressions impose a linear relationship between *Evaluation* and the explanatory variables that may be restrictive, but as a result, the cross term coefficient is the same as the interaction effect of the two variables in *Evaluation*, as long as we confine our attention to the interior between the two endpoints. Thus, tobit models can provide a quick check of the directions of the interaction effects.

The tobit results in Table 4 provide stronger support for our prediction. The coefficient of the interactive term between *Junior to experienced rep ratio* and *Profit* is negative and significant at the 5 percent level even for Model 3 and at the 1 percent level for the rest.

[25] The lower limit is 1 and the upper limit is 5.

The strong statistical significance of the coefficient in Models 4 and 5 is especially notable. The coefficient for the interaction between *Corporate customer share* and *Profit* is negative and statistical significance improved for all the models except for Models 4 and 5. The insignificant results for Models 4 and 5 may simply suggest that the importance of the relationship with corporate customers does not change for each branch year by year as measured by *Corporate customer share*.

Let us consider the following two alternative interpretations. First, our results might simply capture the possible across-worker heterogeneity in the cost of sales effort since the optimal piece rate on sales is a decreasing function of the cost of effort. If the cost of sales effort is higher for young and less experienced workers, a higher *Junior to experienced rep ratio* would lead to a lower average piece rate offered at the branch. This may induce the interaction term between *Junior to experienced rep ratio* and *Profit* to be negative at the aggregated level. In order to distinguish this effect from the one caused by the multitasking issue, we included the interaction term between worker's experience and *Profit* in all the models in unreported regressions. Reassuringly, our results were almost unaffected, rejecting this interpretation.

Second, suppose that workers are competing for the most profitable customers. Then, if there are more junior workers on the team, it is easier for more experienced workers to attract the most profitable customers at the cost of junior workers. In order to reduce this type of behavior, the supervisors might reduce the incentive on sales, leading to the negative interaction effect between *Junior to experienced rep ratio* and *Profit*. However, this explanation is also not likely at Auto Japan, where each worker is given his or her territory to do sales. Workers do not compete with each other for the same customers under this sale territory system.

Thus, the hypothesis that supervisors take into account hard-to-measure tasks such as mentoring and long-term customer relationship building in their evaluations has been supported by the data.

## 5.2 Does the performance evaluation contain information about future sales performance?

Hard-to-measure tasks such as developing good relationships with corporate customers tend to have an impact on future sales performance, as posited pictorially in Figure 1. It is now useful to examine the claim that an evaluation does in fact contain information predicting future sales performance. If this claim is confirmed, the empirical evidence for the hypothesis that managers use subjective performance evaluation to mitigate the multitasking problem will be further reinforced.

Consider the following model that regresses the one-period-ahead Profit on the current *Evaluation* and the current *Profit*. $X_{it}$ contains other variables that are observable to the third party.

$$\begin{aligned} (Profit)_{i,t+1} &= \beta_0 + \beta_1 (Profit)_{it} + \beta_2 (Evaluation)_{it} + \beta' X_{it} \\ &+ (Branch\ fixed\ effects) + u_{it} \end{aligned}$$

If the current evaluation is solely determined by observable performance measures and observable worker characteristics, it should not affect future performance once the current observables are all controlled for. However, if the evaluation contains information on expected future performance as well as observed performance, it should explain future sales performance. This idea parallels that of Hayes and Schaefer (2000), who tested whether the annual performance evaluations of CEOs contain information that predicts firms' future performance.

Table 5 shows the results. In the model, we include worker, supervisor and branch characteristics that may predict future profit. Some branch characteristics, such as location, affect sales, and these time-invariant characteristics should be controlled for. Therefore, we include branch fixed effects. The coefficient for *Evaluation* is positive and significant at the 5 percent level. This confirms that annual performance evaluations capture information that is not fully captured by current profits.

### 5.3 How are the unexplained variations in evaluation associated with worker quits?

We now examine the effect of biases in subjective performance evaluation on worker quits. Let us first discuss the definition of bias in subjective evaluation. Prendergast and Topel (1996) modeled favoritism in terms of supervisors' altruism towards particular workers. In their model, the supervisor's utility depends on the subordinate's wage. Given that the supervisors' reports affect the subordinates' wages, the supervisors overstate the performance of their favorite subordinates and understate the performance of subordinates they dislike in their evaluations.

In MacLeod (2003), the difference between the supervisor's assessment of the worker's performance and the worker's own assessment is labeled as the worker's perceived bias. Unlike taste-based bias as modelled in Prendergast and Topel (1996), perceived bias arises even when both the supervisor and the worker impartially report their beliefs about the worker's performance because the disagreement comes from different priors or non-overlapping information they receive. Nevertheless, in either type of bias, conflict arises and the worker may quit the firm.

To obtain evidence suggesting the effect of such biases on worker quits, we put forward two key variables. First, we constructed an indicator of an "unexplained evaluation gap" defined as a substantially positive or negative residual from our estimated evaluation equation. This variable could contain performance on hard-to-measure tasks that are unobservable to researchers as well as taste-based bias and perceived bias. More formally, we define the residuals of our ordered probit evaluation regressions as:

$$(Residual)_{it} = (Actual\ evaluation)_{it} - (Predicted\ evaluation)_{it}$$
$$where \quad (Predicted\ evaluation)_{it} = \sum_{k=1}^{5} Prob(Evaluation_{it} = k)k \tag{4}$$

Due to discrepancies between actual and predicted evaluations that arise naturally due to the discrete nature of evaluation, any residuals smaller than 0.5 will not cause a shift in the actual evaluation. Thus, we constructed indicators of negative and positive unexplained evaluation gaps that take the value 1 if a residual is large enough to lower (or raise) the evaluation by one level: $I\{\text{Residual}_{it}<-0.5\}$ and $I\{\text{Residual}_{it}>+0.5\}$, where $I\{\}$ is the indicator variable.

We use the evaluation regression in Model 4 (in Table 3), which includes branch fixed effects, to compute residuals. Given that workers are rarely transferred to other branches, if their perception of fairness in evaluations is formed through past experience, branch-specific, time-invariant factors that are taken for granted should be filtered out in calculating bias. This is why we use this model specification. Our negative and positive evaluation gaps correspond to the $12^{th}$ percentile and the $88^{th}$ percentile of the distribution of the residuals, respectively.

One problem with using this variable as a proxy for evaluation bias is that, as we have already discussed, the residuals of the evaluation regressions could reflect any of the following three factors: (1) taste-based evaluation bias, (2) perceived-bias, or (3) the worker's performance on hard-to-measure tasks that are unobservable to researchers. The presence of the third factor obviously causes a problem in isolating the effects of evaluation bias (the first or second factors) on worker quits. For example, a worker who lacks commitment to work may receive a low evaluation due to his poor performance on hard-to-measure tasks and at the same time he is likely to quit his job for reasons unrelated to his evaluation, leading to an overestimation of the effect of evaluation bias on worker quits.[26]

The third factor can be time-varying as well. For example, a manager may stop assigning mentoring tasks to a worker because he does not sufficiently take care of junior representatives. This may make the worker happier because he can earn a higher commission, or this may make him unhappy if the manager gives him a lower evaluation grade than what his sales performance indicates. Thus, a change in task assignment may affect both the worker's evaluation and his decision to quit.

If there is no time-varying component in the third factor, we could correct for endogeneity bias by simply including worker fixed effects in the following linear probability model:

$$Quit_{it} = \beta_1 I\{Residual_{it} < -0.5\} + \beta_2 I\{Residual_{it} > +0.5\} + \beta' Z_{it} + a_i + e_{it} \qquad (5)$$

where $Quit_{it}$ is a dummy variable that takes the value 1 if the $i^{th}$ worker quits at the end of financial year t, or in the middle of financial year t+1. The term, $a_i$, is the worker level fixed effects. However, if the third factor contains a time-varying component, this model will still be unable to distinguish the effects of evaluation bias from the effects of the third factor.

Nevertheless, we can reduce this endogeneity bias by using our second key variable that exploits potentially large changes in evaluation bias at the time of supervisor changes. When a supervisor changes, evaluation bias specific to the supervisor-worker pair should change discontinuously. On the other hand, unobserved performance on hard-to-measure tasks (the third factor) may not change discontinuously unless the task assignment changes. This suggests the use of within-worker changes in the supervisor-worker match fixed effects to reduce the endogeneity bias. Provided that the change in the third factor is small relative

[26] The presence of the third factor could lead to underestimation as well. For example, a worker with strong social skills may spend more time developing customer relationships that lead to a better evaluation. Such a worker would have greater employment opportunities elsewhere and therefore has a higher quit probability, leading to an underestimation.

to the change in evaluation bias at the time of supervisor change, this variable should largely pick up the effects of evaluation bias on worker quits.[27]

To construct the second key variable–the change in supervisor-worker match fixed effects, we re-estimate the evaluation regression using OLS including the same control variables as Models 2, 4 and 5 of Table 3. We compute the change in supervisor-worker match fixed effects only for workers who experienced a change in supervisor. Among the 686 sales representatives in our sample, 409 experienced at least one change in supervisor. If a worker experienced a change in supervisor more than once, the first supervisor was used as the reference to compute the change. The actual variable is constructed as:

$$\begin{aligned} & \Delta(Match\ fixed\ effects_{it}) \\ = \ & (Match\ fixed\ effect_{it}) - (Match\ fixed\ effect_{iT_i}) \end{aligned}$$

where $T_i$ indicates the initial period in which the $i^{th}$ individual appears in the sample. We then estimate the following linear probability model.

$$Quit_{it} = \beta_1 \Delta(Match\ fixed\ effects_{it}) + \beta' Z_{it} + e_{it} \tag{6}$$

One advantage of estimating the two different models (Equations 5 and 6) is that a comparison of the results allows us to infer the magnitude of the endogeneity bias. Note that the use of the second key variable ($\Delta$*Match-fixed effects*) should reduce the endogeneity bias that is expected in the analysis using our first variable (unexplained evaluation gap), where the endogeneity bias arises from time-varying unobserved performance. If the use of the second key variable yields much smaller impact of the unexplained variation in evaluation on worker quits than the use of the first key variable, we might conclude that the third factor is driving the results. In contrast, if the second key variable produces a similar result, then

[27] Workers rarely transfer across branches at Auto Japan. Thus, we cannot estimate the supervisor-worker match fixed effects separately from branch fixed effects. However, by using the change in the estimated match fixed effect, we can eliminate branch fixed effects.

the estimated effect of unexplained variation in the evaluation is likely to come largely from evaluation bias.

**Quit regression results**

In our sample, 75 workers left the company. Our data contain detailed reasons for each separation. Among the 75 workers, 49 took other jobs, 12 left to inherit family businesses, 12 retired, and 2 were fired. We exclude retirements and firings from our definition of quits, so the total number of quits in our sample is 61.

Table 6 shows the estimation results using the "unobserved evaluation gap". Column 1 is a simple probit model that does not account for worker fixed effects. It suggests that workers who received evaluation ratings lower than what is explained by objective performance measures are more likely to quit. The positive evaluation gap has no effect on worker quits. The results are robust to the change in model specification to a linear probability model with worker fixed effects (Column 2), which corrects for the endogeneity bias caused by time-invariant unobservables. According to Column 2 result, the negative evaluation gap would increase the probability of quitting by 4.7 percentage points.

Column 2 does not distinguish between the impact of the unexplained evaluation gap and that of the evaluation itself–when a worker receives a bad evaluation, he might quit not because the bad evaluation is unjustifiable by objective performance measures, but simply because he received a bad evaluation. In order to identify the pure impact of the unexplained evaluation gap, we thus add in Column 3 the explained component of evaluation, the predicted value of evaluation defined by equation 4 rounded to the nearest integer as categorical dummies.

The coefficient for the negative evaluation gap does not become smaller at all, and remains significant. A negative evaluation gap increases quit probability by 5.7 percentage points. The coefficients for the predicted evaluations monotonically increases with evaluation grade in absolute values, showing that a lower predicted evaluation increases quit probabilities.

If the negative evaluation gap simply captures the worker's performance on hard-to-measure tasks that are unobservable to the researchers, the effect of a negative unexplained evaluation gap should be similar to the effect of a one-point decline in the predicted evaluation. On average, a one-point decline in the predicted evaluation increases quit probability by 3.1 percentage points, which is statistically significantly lower than the effect of the negative evaluation gap (p-value=0.016).[28] The larger impact of the negative evaluation gap compared to that of the predicted evaluation indicates that something other than unobservables, such as evaluation bias, might have played some role in inducing a higher quit rate.

Now, we attempt to further reduce the endogeneity bias by using the change in supervisor-worker match fixed effects (Equation 6). Column 1 of Table 7 shows the estimation results with bootstrapped standard errors. The coefficient for $\Delta(Match\ fixed\ effects_{it})$ is negative and significant at the 1 percent level, indicating that a reduction in match fixed effect increases worker quits. Column 2 splits the variable into positive and negative parts to capture the possible non-linear effects. As such, we find a significant effect when there is a reduction in the supervisor-worker match fixed effect, but we do not find a significant effect when there is an increase in the match fixed effect.

What are the magnitudes of these estimated effects? To be comparable with the unobserved evaluation gap analyses in Table 6, we first compute the average of $\Delta$*Match fixed effect* for those who experienced changes that are large enough to downgrade their evaluation by one (i.e., a change in predicted evaluation that is lower than -0.5). The average is

[28]The average effect of a one-point decline in the predicted evaluation is given by $[(\beta_{EvalS}$-$\beta_{EvalA})$+$(\beta_{EvalA}$-$\beta_{EvalB})$ +$(\beta_{EvalB}$-$\beta_{EvalC})$+$(\beta_{EvalC})]/4$=$\beta_{EvalS}/4$.

-0.78. Thus, Column 2 results indicate that if the match fixed effect drops by this much, the quit probability would increase on average by -0.070×-0.78≃ 5.5 percentage points. The magnitude of the effect is similar to the coefficients for the negative evaluation gaps in Column 3 of Table 6.

Note that this analysis was intended to reduce a potential endogeneity bias that might have contaminated the results for the models with the unobserved evaluation gap. The estimated impact that is comparable with our former model may indicate that the estimated effect of unexplained variation in the evaluation on worker quits should largely come from evaluation bias.

Finally, we provide further suggestive evidence that evaluation bias as reflected in the unobserved negative evaluation gap causes worker dissatisfaction. In Appendix A, we examine whether an unexplained evaluation gap predicts worker dissatisfaction with evaluation results using additional information from a survey of workers we conducted. The survey includes a question about how fair they felt the evaluation results were. If the negative evaluation gap only captures poor performance on hard-to-measure tasks, it should not necessarily affect workers' opinions of the fairness of evaluations. Thus, if the negative unexplained evaluation gap is significantly associated with the worker's perceived fairness of the evaluation results, it is consistent with our interpretation that this measure at least partly reflects either taste-based bias or perceived bias.

In Table A.1 of the appendix, we show that a negative evaluation gap indeed reduces workers' perception of fairness about the evaluation results. The impact of a negative evaluation gap on the reported perceived fairness rating (in percentage) is a 13 to 17 percentage point reduction.

# 6. Discussion and Conclusion

Subjective performance evaluation used in an incentive contract cuts two ways. It can mitigate multitasking agency problems, but it also opens the door for biases, resulting in higher turnover and lower job satisfaction. We have provided new evidence for both sides of subjective performance evaluation. Let us briefly summarize our new contributions to the related literature.

First, we showed that subjective evaluation is indeed used to incorporate hard-to-measure tasks in an incentive contract, namely mentoring junior workers and building long-term customer relationships. Specifically, by using proxies for the marginal productivity of these hard-to-measure tasks, we showed that the sensitivity of wages to sales performance declines with these proxies. One of our contributions to the literature is that we looked inside the "black box" and examined what hard-to-measure tasks are actually measured by subjective evaluations. In earlier studies with similar goals, what subjective evaluations are meant to assess is somewhat unclear.

Thus, our results serve as useful complementary evidence for these studies. For example, Gibbs et al. (2004) showed that the amount of spending on personal training is positively related to the use of subjectively determined bonuses. This observation is consistent with our finding that mentoring tasks are rewarded in subjective evaluations. Bushman, Indjejikian, and Smith (1996) found that growth opportunity and the length of the product development cycle are positively related to the use of subjective individual performance evaluations in determining CEO compensation. This result is consistent with our finding that subjective evaluation is used to reward long-term, value enhancing activities.

Our study also provides an answer to the following question raised by Griffith and Neely (2009); "Are managers able to interpret and react to multiple performance indicators?" This is an important question since the effectiveness of multidimensional incentive schemes, such as the balanced score card incentives, depends on the ability of managers to react. Our study shows that the supervisors in Auto Japan were able to respond to local information, such as the importance of mentoring or relationship development with corporate customers, by adjusting the weights for different performance measures presumably to induce appropriate allocation of efforts.[29]

Second, we provided circumstantial yet consistent evidence that negative bias in subjective performance evaluation increases worker quits.[30] Throughout our analyses, we have paid attention to the possibility that worker performance on hard-to-measure tasks that are unobservable to researchers causes a problem in identifying the effects of evaluation bias on worker quits.

In order to deal with this issue, we included the residual-based 'negative evaluation gap' as an explanatory variable in a linear quit probability model with worker fixed effects. To further reduce the endogeneity bias stemming from time-varying unobservables, we estimated the effects of within-worker changes in the supervisor-worker match fixed effects on worker quits. The results from both estimates imply that an unexplained decline in evaluation by one grade would increase quit probability by roughly 5 percentage points. The estimated impact that is robust to model changes, which is intended to reduce endogeneity biases, seems to suggest that taste-based bias or perceived bias play a substantial role in inducing worker quits. We also showed that the negative evaluation gap is negatively asso-

[29] Ittner, Larcker and Meyer (2003) provided an example where balanced scorecard was not effective. They showed that the subjectivity in the scorecard plan in a US retail bank allowed supervisors to ignore qualitative measures that were predictive of future financial performance and award bonuses primarily based on current financial performance. We note however that the results do not necessarily imply that the bank's subjective performance assessments did not help to mitigate multitasking and gaming problems. Other factors considered in the performance evaluation system may be affecting the size of bonuses in non-linear way (e.g., only when an individual score falls short of a certain threshold).

[30] There is some anecdotal evidence for this relationship. Stewart (1993) describes a case in First Boston, where the firm announced that senior managers would receive reduced bonuses. Many senior managers claimed that they had been promised more while the company argued that the bonuses merely reflected disappointing financial results. The dispute ended with many managers leaving the firm. Endlich (1999) also describes a similar a case at Goldman Sachs.

ciated with workers' views on the fairness of their evaluation. Putting this all together, our results are consistent with the hypothesis that evaluation biases lead to higher quit rates.

There are, however, two additional important issues that we have been unable to study using our data set. First, one important task for sales representatives is to "cross-sell" by soliciting repair and maintenance work for their branch's service department, work that can be more profitable than new car sales. In some cases, it is better to divert the sales representatives' efforts towards cross-selling, especially when the capacity utilization of the service department in their branch is low. Therefore, if we had detailed data on cross-selling, such as how many previous customers come back for repair and maintenance work, and data about service departments' capacity utilization (e.g., revenue per mechanic), we could examine in detail whether a high level of cross-selling during a time of low capacity utilization is highly evaluated in subjective performance evaluations.

Second, subjective evaluation should reflect employer learning. Because sales performance is affected by factors beyond the control of sales representatives, supervisors may try to smooth their evaluation grades by taking the weighted average of current performance and the workers' expected ability or productivity. The latter is actually the average of their past evaluation scores. In that case, the weight on current performance should be a decreasing function of how long the worker and supervisor have held their respective posts. Since we only have four years of data, we cannot satisfactorily evaluate this hypothesis. These issues are left for future research.

# References


Ai, Chunrong; and Edward. C. Norton. (2003) "Interaction Terms in Logit and Probit Models." *Economics Letters*, Vol. 80(1), pp. 123-129.

Baker, George. (2002) "Distortion and Risk in Optimal Incentive Contracts." *Journal of Human Resources*, Vol. 37(4), pp. 728-751.

Baker, George; Robert Gibbons, and Kevin J. Murphy. (1994) "Subjective Performance Measures in Optimal Incentive Contracts." *Quarterly Journal of Economics*, Vol.109(4), pp.1125-1156.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. (2005) "Social Preferences and the Response to Incentives: Evidence from Personnel Data." *Quarterly Journal of Economics*, Vol.120(3), pp.917-962.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. (2007) "Incentives for Managers and Inequality among Workers: Evidence from a Firm-level Experiment." *Quarterly Journal of Economics*, Vol.122(2), pp.729-773.

Bartel, Ann; Brianna Cardi and Kathryn Shaw. (2012) "Incentives for Leadership: Multitasking in a Professional Services Firm" Working Paper.

Bull, Clive. (1987) "The Existence of Self-Enforcing Implicit Contracts." *Quarterly Journal of Economics*, Vol. 102(1), pp. 147-159.

Bushman, Robert M., Raffi J. Indjejikian, and Abbie Smith. (1996) "CEO Compensation: The Role of Individual Performance Evaluation." *Journal of Accounting and Economics*, Vol. 21(2), pp. 161-193.

Chan, Jimmy and Bingyong Zheng. (2011) "Rewarding Improvements: Optimal Dynamic Contracts with Subjective Evaluation." *The RAND Journal of Economics*, Vol. 42(4), pp. 758-775.

Drago, Robert and Gerald T. Garvey. (1998) "Incentives for Helping on the Job: Theory and Evidence." *Journal of Labor Economics*, Vol.16(1), pp.1-25.

Dumont, Etienne; Bernard Fortin, Nicolas Jacquemet and Bruce Shearer. (2008) "Physicians' Multitasking and Incentives: Empirical Evidence from a Natural Experiment." *Journal of Health Economics*, Vol. 27(6), pp.1436-1450.

Ederhof, Merle. (2010) "Discretion in Bonus Plans." *Accounting Review*, Vol. 85(6), pp. 1921-1949.

Elvira, Marta and Robert Town. (2001) "The Effects of Race and Worker Productivity on Performance Evaluations." *Industrial Relations*, Vol. 40(4), pp. 571-590.

Endlich, Lisa. (1999) *Goldman Sachs: The Culture of Success*, New York: Alfred A. Knopf.

Engellandt, Axel and Regina T. Riphahn. (2011) "Evidence on Incentive Effects of Subjective Performance Evaluations." *Industrial and Labor Relations Review*, Vol.64(2), pp.241-257.

Fuchs, William. (2007) "Contracting with Repeated Moral Hazard and Private Evaluations." *American Economic Review*, Vol. 97(4), pp. 1432-1448.

Gibbs, Michael, Kenneth A. Merchant, and Wim A. Van der Stede and Vargus, Mark E. (2004) "Determinants and Effects of Subjectivity in Incentives." *Accounting Review*, Vol.79(2), pp. 409-436.

Giuliano, Laura, David I. Levine, and Jonathan Leonard. (2005): "Do Race, Gender, and Age Differences Affect Manager-Employee Relations? An Analysis of Quits, Dismissals, and Promotions at a Large Retail Firm." mimeo, University of California, Berkeley.

Goldin, Claudia and Cecilia Rouse. (2000) "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians." *American Economic Review*, Vol. 90(4), pp. 715-741.

Griffith, Rachel and Andrew Neely. (2009) "Performance Pay and Managerial Experience in Multitask Teams: Evidence from within a Firm." *Journal of Labor Economics*, Vol. 27(1), Jan., pp. 49-82.

Haley, M. Ryan. (2003). "The Response of Worker Effort to Piece Rates: Evidence from the Midwest Logging Industry." *Journal of Human Resources*, Vol.38(4), pp.881-890.

Hayes, Rachel M. and Scott Schaefer. (2000) "Implicit Contracts and the Explanatory Power of Top Executive Compensation for Future Performance." *The RAND Journal of Economics*, Vol. 31(2), pp. 273-293.

Healy, Paul M. (1985) "The Effect of Bonus Schemes on Accounting Decisions." *Journal of Accounting and Economics*, Vol. 7(1-3), pp.85-107.

Holmstrom, Bengt and Paul Milgrom. (1991) "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics and Organization*, Vol. 7(2), pp. 24-52.

Holthausen, Robert W., David F. Larcker, and Richard G. Sloan. (1995) "Business unit innovation and the structure of executive compensation." *Journal of Accounting and Economics*, Vol. 19(2/3), pp.279-313.

Hoppe, Felix. and Frank Moers. (2011) "The Choice of Different Types of Subjectivity in CEO Annual Bonus Contracts." *Accounting Review*, Vol.86 (6), pp. 2023-2046.

Ittner, Christopher D., David F. Larcker and Marshall W. Meyer. (2003) "Subjectivity and the Weighting of Performance Measures: Evidence from a Balanced Scorecard." *Accounting Review*, Vol. 78(3), pp. 725-758.

Kerr, Steven. (1975) "On the Folly of Rewarding A While Hoping for B." *Academy of Management Journal*, Vol.18(4), pp.769-783.

Larkey, Patrick and Jonathon Caulkins. (1992) "All Above Average." mimeo, Carnegie Mellon University.

Larkin, Ian. (2014) "The Cost of High-Powered Incentives: Employee Gaming in Enterprise Software Sales." *Journal of Labor Economics*, Vol. 32(2), pp. 199-227

Laurence, J. Peter and Raymond Hull. (1969) *The Peter Principle: Why Things Always Go Wrong*, William Morrow and Company, Inc., New York.

Lazear, Edward P. (2000) "Performance Pay and Productivity." *American Economic Review*, Vol. 90(5), pp. 1346-1361.

Levin, Jonathan. (2003) "Relational Incentive Contracts." *American Economic Review*, Vol. 93(3), pp. 835-857.

MacLeod, W. Bentley. (2003) "Optimal Contracting with Subjective Evaluation." *American Economic Review*, Vol. 93(1), pp. 216-240.

MacLeod, W. Bentley and James M. Malcomson. (1989) "Implicit Contracts, Incentive Compatibility, and Involuntary Unemployment." *Econometrica*, Vol. 57(2), pp. 447-480.

Mannix, Elizabeth and Margaret A Neale. (2005) "What Differences Make a Difference?: the Promise and Reality of Diverse Teams in Organizations." *Psychological Science in the Public Interest*, Vol. 6(2) (October): 31-55.

Merchant, Kenneth. A. (1989) *Rewarding Results: Motivating Profit Center Managers*, Boston, MA: Harvard Business School Press.

Murphy, Kevin. J. and Paul Oyer. (2003) "Discretion in Executive Incentive Contracts." Working Paper, University of Southern California and Stanford University.

Murphy, Kevin R. and Jeanette Cleveland. (1991) *Performance Appraisal: An Organizational Perspective*, Boston: Allyn and Bacon.

Owan, Hideo and Tsuyoshi Tsuru. (2011) "Integrating High-Powered Performance Pay into a Seniority Wage System." mimeo, Hitotsubashi University.

Oyer, Paul. (1998) "Fiscal Year Ends and Nonlinear Incentive Contracts: the Effect on Business Seasonality." *Quarterly Journal of Economics*, Vol.113(1), pp.149-185.

Paarsch, Harry J. and Bruce S Shearer. (1999) "The Response of Worker Effort to Piece Rates: Evidence from the British Columbia Tree-Planting Industry." *Journal of Human Resources*, Vol.34(4), pp.643-667.

Paarsch, Harry J. and Bruce S Shearer. (2000) "Piece Rates, Fixed Wages and Incentive Effects: Statistical Evidence from Payroll Records" *International Economic Review*, Vol.41(1), pp.59-92.

Pearce, David G. and Stacchetti, Ennio. (1998) "The Interaction of Implicit and Explicit Contracts in Repeated Agency." *Games and Economic Behavior*, Vol. 23(1), pp.75-96.

Prendergast, Canice. (1999) "The Provision of Incentives in Firms." *Journal of Economic Literature*, Vol. 37(1), pp. 7-63.

Prendergast, Canice and Robert H. Topel. (1993) "Discretion and Bias in Performance Evaluation." *European Economic Review*, Vol.37(2/3), pp. 355-365.

Prendergast, Canice and Robert H. Topel. (1996) "Favoritism in Organizations." *Journal of Political Economy*, Vol.104(5), pp.958-978.

Stewart, James B. (1993) "Taking the Dare." *The New Yorker*, July 26, pp. 34-39.

Thiele, Veikko. (2013) "Subjective Performance Evaluations, Collusion, and Organizational Design, Journal of Law, Economics, & Organization, 29(1), 35-59.

Zenger, Todd R., and Barbara S. Lawrence. (1989). "Organizational Demography: The Differential Effects of Age and Tenure Distributions on Technical Communication." *Academy of Management Journal*, Vol. 32(2), pp. 353-376.

Table 1: Wage raise matrix (Raise in monthly base salary in 1000 yen)

| | | Evaluation results | | | | |
|---|---|---|---|---|---|---|
| | | s | a | b | c | d |
| Salary | S | 2.3 | -2.2 | -40 | -50 | -70 |
| Stage | A | 5.6 | 2.1 | -2.1 | -40 | -50 |
| | B | 40 | 5.2 | 2 | -1.9 | -40 |
| | C | 50 | 40 | 4.8 | 1.8 | -1.8 |
| | D | 70 | 50 | 40 | 4.4 | 1.7 |

Figure 1: Hard to measure tasks

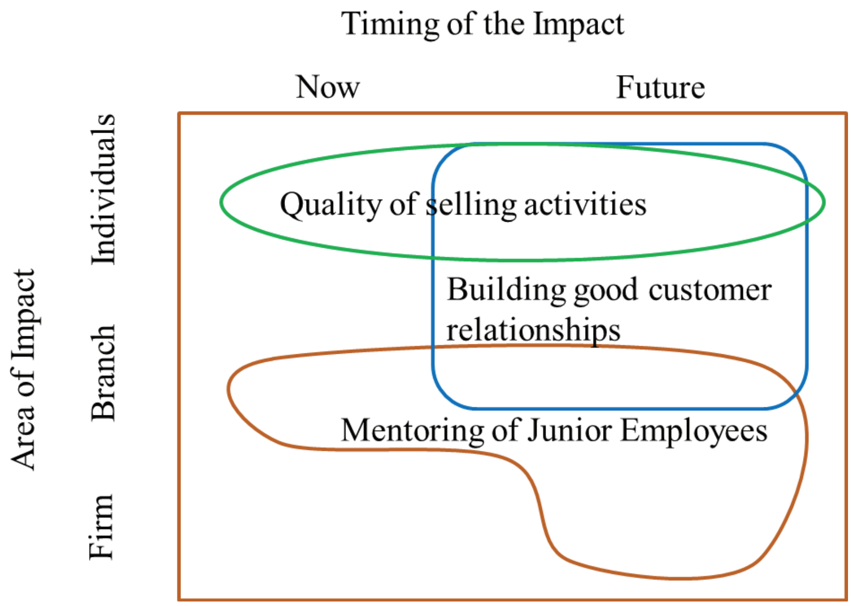

Table 2: Summary Statistics: # Obs=2148

| Variables | Mean | St Dev | Min | Max |
|---|---|---|---|---|
| Evaluation | 2.664 | (0.837) | 1 | 5 |
| Evaluation=s | 0.008 | (0.091) | 0 | 1 |
| Evaluation=a | 0.141 | (0.348) | 0 | 1 |
| Evaluation=b | 0.430 | (0.495) | 0 | 1 |
| Evaluation=c | 0.348 | (0.477) | 0 | 1 |
| Evaluation=d | 0.073 | (0.260) | 0 | 1 |
| Profit (million yen) | 20.716 | (7.991) | 1.373 | 61.578 |
| Junior to experienced rep ratio | 0.046 | (0.101) | 0 | 0.667 |
| Corporate customer share | 0.267 | (0.105) | 0.076 | 0.796 |
| Branch-firm productivity difference (million yen)×(it is positive) | 1.224 | (2.130) | 0 | 14.370 |
| Branch-firm productivity difference (million yen) ×(it is negative) | -1.149 | (1.585) | -7.960 | 0 |
| Worker's tenure | 12.876 | (9.518) | 0.750 | 42 |
| Supervisor's tenure | 25.753 | (5.712) | 5 | 36 |
| Worker's educ=university | 0.650 | (0.477) | 0 | 1 |
| Worker's educ=vocational school | 0.123 | (0.329) | 0 | 1 |
| Worker's educ=high school | 0.214 | (0.410) | 0 | 1 |
| (The excluded category=below high school) | | | | |
| #Sales reps at the branch[(a)] | 8.273 | (2.033) | 3.833 | 13 |
| Average worker tenure at the barnch | 12.808 | (3.216) | 5.329 | 22.561 |
| Sd of worker tenure at the branch | 9.227 | (2.765) | 2.871 | 18.507 |
| Salary stage=S | 0.027 | (0.161) | 0 | 1 |
| Salary stage=A | 0.101 | (0.302) | 0 | 1 |
| Salary stage=B | 0.259 | (0.438) | 0 | 1 |
| Salary stage=C | 0.307 | (0.461) | 0 | 1 |
| Salary stage=D | 0.306 | (0.461) | 0 | 1 |
| Year 2001 | 0.256 | (0.436) | 0 | 1 |
| Year 2002 | 0.251 | (0.434) | 0 | 1 |
| Year 2003 | 0.242 | (0.428) | 0 | 1 |

(a) When a worker workers for a fraction of a year, the fraction is added in the computation of # Sales reps at the branch. Thus, this variable can take a non-integer value.

Figure 2: Evaluation and profits from car sales

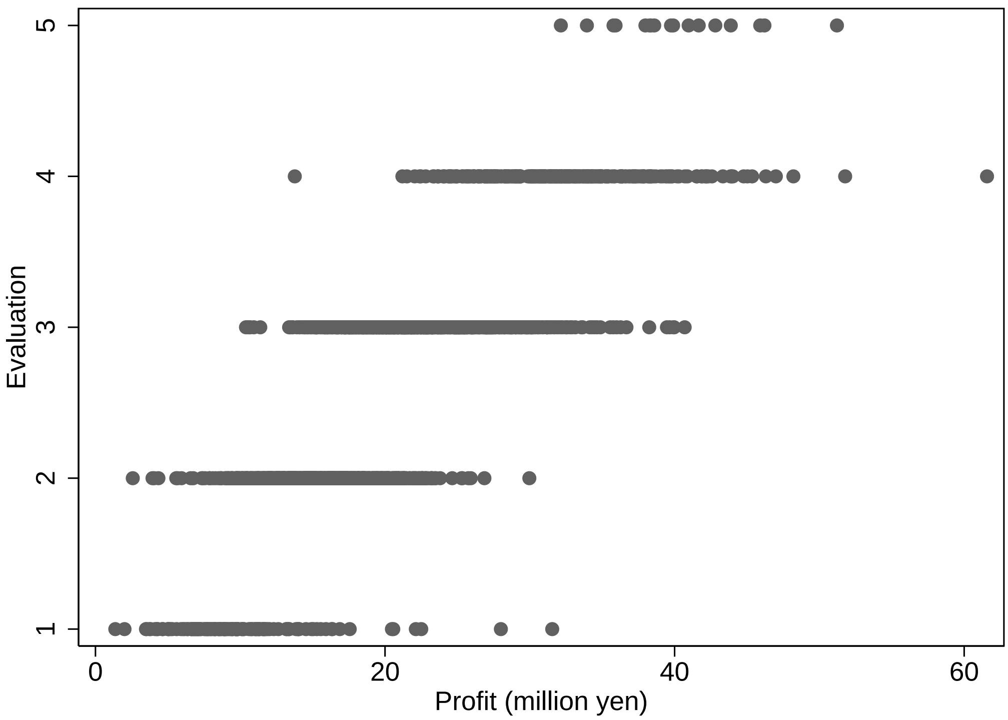

Table 3: Ordered probit evaluation regressions

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Profit | 0.261 *** | 0.263 *** | 0.248 *** | 0.274 *** | 0.282 *** |
| | (0.013) | (0.014) | (0.013) | (0.014) | (0.015) |
| (Profit)×(Junior to experienced rep ratio) | -0.107 ** | -0.081 * | -0.071 | -0.101 ** | -0.080 |
| | (0.045) | (0.048) | (0.048) | (0.049) | (0.051) |
| Junior to experienced rep ratio | 2.382 ** | 2.162 * | 1.766 | 3.013 ** | 2.474 |
| | (1.022) | (1.138) | (1.172) | (1.193) | (1.253) |
| (Profit)×(Corporate customer share) | -0.103 *** | -0.090 ** | -0.104 *** | -0.085 * | -0.081 * |
| | (0.038) | (0.040) | (0.036) | (0.044) | (0.047) |
| Corporate customer share | 2.650 *** | 2.240 *** | 2.554 *** | 1.637 | 1.287 |
| | (0.693) | (0.764) | (0.728) | (1.169) | (1.257) |
| Branch-firm productivity difference | | -0.033 * | -0.026 | -0.066 ** | -0.042 * |
| ×(it is positive) | | (0.020) | (0.018) | (0.029) | (0.024) |
| Branch-firm productivity difference | | -0.088 *** | -0.081 *** | -0.162 *** | -0.129 *** |
| ×(it is negative) | | (0.022) | (0.022) | (0.034) | (0.035) |
| Worker's tenure | | 0.086 *** | 0.038 *** | 0.087 *** | 0.090 *** |
| | | (0.014) | (0.015) | (0.015) | (0.015) |
| Worker's tenure$^2$ | | -0.001 *** | -0.001 *** | -0.001 *** | -0.001 *** |
| | | (0.000) | (0.000) | (0.000) | (0.000) |
| Supervisor's tenure | | 0.001 | -0.002 | -0.010 * | -2.277 |
| | | (0.007) | (0.007) | (0.006) | (1.418) |
| Worker's educ=university | | 0.801 | 0.703 | 0.944 ** | 1.026 ** |
| | | (0.499) | (0.548) | (0.474) | (0.492) |
| Worker's educ=vocational school | | 0.660 | 0.568 | 0.846 * | 0.954 * |
| | | (0.514) | (0.558) | (0.495) | (0.513) |
| Worker's educ=high school | | 0.834 * | 0.732 | 0.968 ** | 1.085 ** |
| | | (0.486) | (0.535) | (0.457) | (0.473) |
| #Sales reps at the branch | | -0.022 | -0.068 | -0.839 * | -0.268 |
| | | (0.104) | (0.102) | (0.434) | (0.277) |
| #Sales reps at the branch$^2$ | | 0.002 | 0.005 | 0.037 | 0.018 |
| | | (0.006) | (0.006) | (0.024) | (0.016) |
| Average worker tenure at the branch | | 0.015 | 0.014 | 0.070 *** | 0.059 * |
| | | (0.012) | (0.012) | (0.023) | (0.023) |
| Sd of worker tenure at the branch | | -0.044 *** | -0.038 ** | -0.088 *** | -0.060 ** |
| | | (0.016) | (0.016) | (0.025) | (0.029) |
| Salary stage=S | | | 2.144 *** | | |
| | | | (0.233) | | |
| Salary stage=A | | | 2.104 *** | | |
| | | | (0.175) | | |
| Salary stage=B | | | 1.533 *** | | |
| | | | (0.125) | | |
| Salary stage=C | | | 0.803 *** | | |
| | | | (0.086) | | |
| Year dummies | Yes | Yes | Yes | Yes | Yes |
| Branch dummies | No | No | No | Yes | No |
| Supervisor dummies | No | No | No | No | Yes |
| Pseudo R squared | 0.45 | 0.50 | 0.53 | 0.53 | 0.54 |
| #Obs | 2148 | 2148 | 2148 | 2148 | 2148 |

Cluster robust sd errors at sales group level are in the parentheses. *, **, ***, significant at 10, 5, 1 percent.

Table 4: Tobit Evaluation Regressions

| Variable | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Profit | 0.103 *** | 0.094 *** | 0.083 *** | 0.093 *** | 0.092 *** |
| | (0.005) | (0.005) | (0.004) | (0.005) | (0.005) |
| (Profit)×(Junior to experienced rep ratio) | -0.067 *** | -0.047 *** | -0.041 ** | -0.051 *** | -0.044 *** |
| | (0.018) | (0.018) | (0.017) | (0.017) | (0.017) |
| Junior to experienced rep ratio | 1.584 *** | 1.274 *** | 1.086 *** | 1.544 *** | 1.404 *** |
| | (0.429) | (0.435) | (0.416) | (0.430) | (0.423) |
| (Profit)×(Corporate customer share) | -0.038 ** | -0.028 * | -0.031 ** | -0.023 | -0.020 |
| | (0.015) | (0.014) | (0.012) | (0.015) | (0.015) |
| Corporate customer share | 1.010 *** | 0.714 ** | 0.796 *** | 0.350 | 0.191 |
| | (0.302) | (0.282) | (0.259) | (0.405) | (0.422) |
| Branch-firm productivity difference | | -0.014 ** | -0.011 * | -0.029 *** | -0.018 ** |
| ×(it is positive) | | (0.007) | (0.006) | (0.010) | (0.008) |
| Branch-firm productivity difference | | -0.037 *** | -0.033 *** | -0.061 *** | -0.046 *** |
| ×(it is negative) | | (0.008) | (0.008) | (0.012) | (0.012) |
| Worker's tenure | | 0.038 *** | 0.018 *** | 0.037 *** | 0.036 *** |
| | | (0.005) | (0.005) | (0.005) | (0.005) |
| Worker's tenure$^2$ | | -0.001 *** | 0.000 *** | -0.001 *** | -0.001 *** |
| | | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| Supervisor's tenure | | 0.000 | -0.001 | -0.003 * | -0.631 |
| | | (0.003) | (0.002) | (0.002) | (0.524) |
| Worker's educ=university | | 0.245 | 0.199 | 0.279 * | 0.278 * |
| | | (0.178) | (0.178) | (0.160) | (0.162) |
| Worker's educ=vocational school | | 0.184 | 0.145 | 0.238 | 0.246 |
| | | (0.183) | (0.181) | (0.166) | (0.168) |
| Worker's educ=high school | | 0.251 | 0.205 | 0.283 * | 0.297 * |
| | | (0.174) | (0.175) | (0.155) | (0.157) |
| #Sales reps at the branch | | -0.017 | -0.032 | -0.303 * | -0.084 |
| | | (0.038) | (0.033) | (0.157) | (0.097) |
| #Sales reps at the branch$^2$ | | 0.001 | 0.002 | 0.013 | 0.006 |
| | | (0.002) | (0.002) | (0.009) | (0.006) |
| Average worker tenure at the branch | | 0.007 | 0.007 | 0.029 *** | 0.026 *** |
| | | (0.005) | (0.004) | (0.008) | (0.008) |
| Sd of worker tenure at the branch | | -0.019 *** | -0.016 *** | -0.034 *** | -0.025 ** |
| | | (0.006) | (0.006) | (0.009) | (0.010) |
| Salary stage=S | | | 0.737 *** | | |
| | | | (0.073) | | |
| Salary stage=A | | | 0.717 *** | | |
| | | | (0.057) | | |
| Salary stage=B | | | 0.528 *** | | |
| | | | (0.041) | | |
| Salary stage=C | | | 0.314 *** | | |
| | | | (0.031) | | |
| Year dummies | Yes | Yes | Yes | Yes | Yes |
| Branch dummies | No | No | No | Yes | No |
| Supervisor dummies | No | No | No | No | Yes |
| Pseudo R squared | 0.40 | 0.46 | 0.49 | 0.48 | 0.50 |
| #Obs | 2148 | 2148 | 2148 | 2148 | 2148 |

Cluster robust sd errors at sales group level are in the parentheses. *, **, ***, significant at 10, 5, 1 percent.

Table 5: Does the performance evaluation contain information about future sales performance?

Dependent Variable=$\text{Profit}_{i,t+1}$

| Variables | Coefficients |
| --- | --- |
| $\text{Evaluation}_{it}$ | 0.580 ** |
| | (0.278) |
| $\text{Profit}_{it}$ | 0.700 *** |
| | (0.031) |
| Worker's $\text{tenure}_{it}$ | -0.103 *** |
| | (0.022) |
| Supervisor's $\text{tenure}_{it}$ | 0.002 |
| | (0.032) |
| Branch-firm productivity $\text{difference}_{it}$ | -0.690 *** |
| $\times$(it is positive) | (0.110) |
| Branch-firm productivity $\text{difference}_{it}$ | -0.755 *** |
| $\times$(it is negative) | (0.134) |
| Average worker tenure at the $\text{branch}_{it}$ | 0.094 |
| | (0.120) |
| Sd of worker tenure at the $\text{branch}_{it}$ | 0.165 |
| | (0.151) |
| #Sales reps at the $\text{branch}_{it}$ | -1.166 |
| | (2.240) |
| #Sales reps at the $\text{branch}_{it}{}^2$ | 0.055 |
| | (0.114) |
| Salary $\text{stage}_{it}$=S | 4.305 ** |
| | (1.841) |
| Salary $\text{stage}_{it}$=A | 2.680 *** |
| | (0.913) |
| Salary $\text{stage}_{it}$=B | 1.703 *** |
| | (0.579) |
| Salary $\text{stage}_{it}$=C | 1.222 *** |
| | (0.429) |
| Constant | 8.033 |
| | (10.954) |
| Year dummies | Yes |
| R squared (within) | 0.63 |
| #Obs | 1393 |

Cluster robust sd errors at the worker level are in the parentheses. *, **, ***, significant at 10, 5, 1 percent.

Table 6: Predicting worker quits using the residuals of the evaluation regression

| | Probit | Linear prob worker fixed effects | |
|---|---|---|---|
| Variables | (1) | (2) | (3) |
| $\text{Residual}_{it}$<-0.5 | 0.755 *** | 0.047 *** | 0.057*** |
| | (0.194) | (0.014) | (0.018) |
| $\text{Residual}_{it}$>-0.5 | 0.031 | 0.007 | -0.003 |
| | (0.246) | (0.009) | (0.011) |
| Profit | -0.115 *** | 0.000 | 0.003 |
| | (0.030) | (0.002) | (0.002) |
| (Profit)×(Junior to Experienced Rep Ratio) | 0.075 | 0.000 | -0.001 |
| | (0.120) | (0.003) | (0.003) |
| Junior to Experienced Rep Ratio | -0.566 | 0.011 | 0.026 |
| | (1.954) | (0.084) | (0.083) |
| (Profit)×(Corporate Customer Share) | 0.154 ** | -0.009 | -0.010 |
| | (0.076) | (0.007) | (0.007) |
| Corporate customer Share | -2.143 | 0.318 | 0.339* |
| | (1.551) | (0.202) | (0.200) |
| (Store-firm productivity difference) | 0.032 | -0.004 | -0.004 |
| ×(it is positive) | (0.050) | (0.003) | (0.003) |
| (Store-firm productivity difference) | 0.010 | -0.002 | -0.002 |
| ×(it is negative) | (0.033) | (0.002) | (0.002) |
| Worker tenure | 0.005 | 0.026 *** | 0.030*** |
| | (0.029) | (0.004) | (0.007) |
| Worker tenure$^2$ | -0.003 | -0.001 *** | -0.001*** |
| | (0.002) | (0.0001) | (0.000) |
| Worker's education in years | -0.110 *** | | |
| | (0.041) | | |
| Supervisor's education in years | 0.042 | | |
| | (0.055) | | |
| Supervisor's tenure | 0.003 | 0.000 | -0.001 |
| | (0.016) | (0.001) | (0.001) |
| Average worker tenure at the store | -0.003 | -0.005 ** | -0.004* |
| | (0.037) | (0.002) | (0.002) |
| S.D. of worker tenure at the store | 0.008 | -0.002 | -0.003 |
| | (0.038) | (0.002) | (0.002) |
| Predicted Eval=S | | | -0.124 |
| | | | (0.086) |
| Predicted Eval=A | | | -0.080 |
| | | | (0.062) |
| Predicted Eval=B | | | -0.039 |
| | | | (0.049) |
| Predicted Eval=C | | | 0.001 |
| | | | (0.035) |
| Constant | 1.141 | -0.113 | -0.197** |
| | (1.214) | (0.082) | (0.082) |
| R squared (Pseud or within) | 0.21 | 0.07 | 0.08 |
| #Obs | 2148 | 2148 | 2148 |

For (1) to (4), bootstrapped sd errors are in the parentheses. For 2SLS model, cluster robust sd errors at the sales group level are in the parentheses. Residuals are computed from Table 3 Model 4. *, **, ***, significant at 10, 5, 1 percent.

Table 7: Predicting worker quits using supervisor-worker match fixed effects. (Dep var = $\text{Quit}_{it}$)

| Variables | (1) | (2) |
|---|---|---|
| $\Delta$ Supervisor-worker match fixed effect | -0.048 *** | |
| | (0.015) | |
| $\Delta$ Supervisor-worker match fixed effect | | -0.070 *** |
| $\times$(it is negative) | | (0.027) |
| $\Delta$ Supervisor-worker match fixed effect | | -0.025 |
| $\times$(it is positive) | | (0.021) |
| Profit | -0.008 *** | -0.008 *** |
| | (0.003) | (0.003) |
| (Profit)$\times$(Junior to experienced rep ratio) | -0.001 | 0.000 |
| | (0.005) | (0.005) |
| Junior to experienced rep ratio | 0.069 | 0.046 |
| | (0.163) | (0.171) |
| (Profit)$\times$(Corporate customer share) | 0.009 | 0.009 |
| | (0.006) | (0.006) |
| Corporate customer share | -0.160 | -0.154 |
| | (0.141) | (0.138) |
| (Branch-firm productivity difference) | 0.002 | 0.002 |
| $\times$(it is positive) | (0.002) | (0.002) |
| (Branch-firm productivity difference) | -0.001 | -0.002 |
| $\times$(it is negative) | (0.004) | (0.004) |
| Worker tenure | -0.003 | -0.003 |
| | (0.003) | (0.003) |
| Worker's tenure$^2$ | 0.000 | 0.000 |
| | (0.000) | (0.000) |
| Worker's education (years) | 0.000 | 0.000 |
| | (0.003) | (0.003) |
| Supervisor's education (years) | 0.001 | 0.001 |
| | (0.002) | (0.002) |
| S.D. of worker tenure at the branch | 0.000 | 0.000 |
| | (0.001) | (0.001) |
| Supervisor's tenure | -0.002 | -0.002 |
| | (0.002) | (0.002) |
| Average worker tenure at the branch | 0.001 | 0.001 |
| | (0.002) | (0.002) |
| Constant | 0.268 | 0.242 |
| | (0.103) | (0.102) |
| Evaluation dummies | Yes | Yes |
| Year dummies | Yes | Yes |
| R squared | 0.08 | 0.08 |
| #Obs | 766 | 766 |

Inside parentheses are bootstrapped standard errors. For the first stage regression to compute the match fixed effects, we used the full 2148 observations (In STATA's bootstrap routine, nodrop option was used.). *, **, ***, significant at 10, 5, 1 percent.

# Appendix A

## Do the bias indicators predict the workers' perceived fairness?

We investigate whether the bias measures we constructed in Section 5.3 in fact predict the workers' acceptance of their evaluation results. In August 2006, we conducted a survey of workers who were randomly sampled at Auto Japan. Among new car sales representatives, the survey was distributed to a sample of 297 subjects, of which 291 responded. Among the respondents, 241 were in our 2003 sample. Thus, our perceived fairness regressions contain only 241 observations. The survey asked the respondents, among other things, to rate the fairness of their evaluations in percentage terms; the higher the percentage, the more fair the evaluation was perceived to be.[31] The survey also asked respondents whether they had received feedback from their supervisors regarding their evaluation results and 59 percent said they had received some feedback. The questionnaires included worker identification numbers, so this survey data could be merged with the sales and evaluation data.

One drawback of our survey was that it was conducted after our sample period. At the time of the survey, the most recent evaluation that workers had received was for the 2005 fiscal year, while our data ends with the 2003 fiscal year evaluation. This gap forces us to use the 2003 explanatory variables to predict perceived fairness in 2005. However, the responses to the survey would reflect the workers' experiences over the intervening period. In addition, a substantial portion of workers were working under the same supervisors in 2005 as in 2003. Thus, 2003 bias indicators would still have some power to predict the levels of perceived fairness reported in 2005.

[31] The respondents were assured that individual responses are used only for academic research and that only the summary information would be given to the management.

Note that workers who moved to other sections, were promoted to supervisor, or quit prior to the survey are not included in the survey. To correct for the potential bias stemming from this sample selection, we include the inverse Mill's ratio in the perceived fairness equation.[32] To separate the effects of bias from low evaluations, the regression models also control for evaluation itself.

Table A.1 shows the tobit regression results with bootstrapped standard errors. Column 1 predicts perceived fairness using our potential bias indicators. The dummy for a negative evaluation gap has a negative and statistically significant coefficient. The computed marginal effect (not shown in the table) indicates that a negative evaluation gap would decrease the perceived fairness rating by 17 percentage points. This model, however, does not control for evaluation itself. To separate the effect of bias from the effect of low evaluation, Column 2 controls for evaluation itself. The coefficient for the negative evaluation gap somewhat reduced, but it is still significant at the 5 percent level. The marginal effect of the negative evaluation gap is a 13 percentage point decrease in the perceived fairness rating.

Column 3 includes the interaction between the negative evaluation gap and feedback to capture the possibility that the effect of bias may differ depending on whether one has received any feedback. As such, there is a negative and statistically significant effect of the negative evaluation gap for those who did not receive any feedback. The computed marginal effect indicates that a negative evaluation gap would decrease the perceived fairness rating by 30 percentage points for those who did not receive any feedback. On the other hand, negative evaluation gap has no effect for those who received some feedback. This result shows the importance of feedback in reducing the perception of unfairness in evaluation.

[32] We derived the inverse Mill's ratio from the probit model where the dependent variable is the dummy variable indicating whether a particular observation is included in the perceived fairness regression. The selection equation contains the same variables in Model 4 (Table 3) except the year and branch dummies.

Table A.1: Determinants of perceived fairness about the evaluation results: Tobit regressions, Dept Var=Perceived Fairness.

| Variables | (1) | (2) | (3) |
|---|---|---|---|
| Residual$_{it}$<-0.5 | -17.766 *** | -13.856 ** | -31.689 ** |
| | (6.825) | (7.025) | (14.816) |
| (Residual$_{it}$<-0.5) $\times$ (Received Feedback) | | | 29.036 * |
| | | | (17.477) |
| Residual$_{it}$>+0.5 | 7.589 | -0.596 | -0.471 |
| | (5.113) | (5.997) | (6.042) |
| Received Feedback | 11.844 *** | 11.538 *** | 8.902 *** |
| | (3.233) | (3.093) | (3.272) |
| Worker tenure | 0.659 | -0.301 | -0.241 |
| | (0.609) | (0.665) | (0.660) |
| Worker tenure$^2$ | -0.008 | 0.014 | 0.013 |
| | (0.017) | (0.019) | (0.018) |
| Worker's education (years) | 0.838 | 1.131 | 1.463 |
| | (1.272) | (1.231) | (1.186) |
| Supervisor's education (years) | 0.004 | -0.048 | -0.002 |
| | (0.366) | (0.360) | (0.378) |
| Supervisor tenure | -1.274 | -1.300 | -1.224 |
| | (0.852) | (0.929) | (0.954) |
| (Branch-firm productivity difference) | 1.156 | 0.649 | 0.462 |
| $\times$(it is positive) | (0.847) | (0.811) | (0.815) |
| (Branch-firm productivity difference) | -0.795 | -1.272 | -1.057 |
| $\times$(it is negative) | (0.987) | (1.010) | (0.998) |
| Average worker tenure at the branch | 0.440 | 0.912 | 0.828 |
| | (0.668) | (0.610) | (0.544) |
| S.D. of worker tenure at the branch | -0.124 | -0.306 | -0.339 |
| | (0.784) | (0.741) | (0.727) |
| Evaluation=s | | 42.543 | 44.928 |
| | | (57.726) | (58.517) |
| Evaluation=a | | 16.238 * | 18.784 ** |
| | | (8.463) | (8.005) |
| Evaluation=b | | 8.665 | 11.122 |
| | | (7.146) | (6.792) |
| Evaluation=c | | -1.312 | 2.386 |
| | | (7.021) | (7.239) |
| Inverse Mill's Ratio | 20.008 * | 7.908 | 8.403 |
| | (12.117) | (10.545) | (10.183) |
| Constant | 22.499 | 36.433 | 28.431 |
| | (35.838) | (34.831) | (33.975) |
| Puseud R squared | 0.02 | 0.02 | 0.03 |
| # Obs | 241 | 241 | 241 |

Bootstrapped sd errors are in the parentheses. For the first stage regression to compute the bias measures, we used the full 2148 observations (In STATA's bootstrap routine, nodrop option was used.). Tobit regressions use zero as the lower bound and 100 as the upper bound. *, **, ***, significant at 10, 5, 1 percent. Residuals are computed from Table 3 Model 4.