ABSTRACT

Essays on Leniency Effects and Narrative Feedbacks in Subjective Performance Evaluation

Jeong-Hoon Hyun
Business School
Seoul National University

This thesis is comprised of two related but independent essays, which examine the incentives of evaluators in subjective performance evaluation. I focus on the process of subjective performance evaluation because we know little about how evaluators give rating and provide narrative feedback, despite practical and academic importance of subjective performance evaluation.

The first essay examines whether supervisors' leniency in subjective performance evaluation is influenced by the prior performance information of subordinates. While prior studies have used proxies for leniency based on aggregate objective performance level or median of subjective scales, I develop the proxy for leniency at the performance measure level based on prior performance level which is found to be more relevant to the current subjective performance score. I analyze archival performance evaluation data of multiple state-owned enterprises (SOEs) in Korea over multiple time periods which enables me to develop the new definition of leniency. Utilizing this data set, I empirically find that lenient rating persists over time, and that larger amount of leniency is applied to low prior performers and harshly evaluated performers in the previous period and smaller amount of leniency is applied to high prior performers and leniently evaluated performers in the previous period. The results have important implications for understanding the incentives of raters which differ with respect to their previous performance information.

The second essay examines the determinants of narrative feedback amounts in subjective performance evaluation. The data of narrative feedbacks in Korean SOEs provides a unique setting to investigate the evaluators' motivation for giving narrative feedback in subjective performance evaluation. I investigate the characteristics of performance measures, evaluators, and evaluatees that affect evaluators' decision for the voluntary feedback about the evaluatees' performance. I find that evaluators present large amount of narrative feedbacks about the highly weighted, long–tenured, and both high scored and low scored performance measures, and SOEs with large number of evaluators and large number of employees. More importantly, I divide narrative feedbacks into those of including suggestions and others, and find that the ratio of narrative suggestions to total narrative comments are associated with performance score, performance measure age and uniqueness, and evaluators' experience, gender and evaluator group size. Overall, the evidence is consistent with evaluators giving substantial narrative feedback when evaluators face low information acquisition cost, high confrontation cost and high intrinsic motivation for providing comments.

Keywords: Performance evaluation, Subjective measure, Prior performance, Leniency effect, Narrative comment, Suggestion for development, Information acquisition cost, Confrontation cost.

Student Number: 2010-30157

TABLE OF CONTENTS

ESSAY1. THE EFFECT OF PRIOR PERFORMANCE INFORMATION ON LENIENCY IN SUBJECTIVE PERFORMANCE EVALUATION

ESSAY2. DETERMINANTS OF NARRATIVE FEEDBACK IN SUBJECTIVE PERFORMANCE EVALUATION

LIST OF FIGURES AND TABLES

ESSAY1. THE EFFECT OF PRIOR PERFORMANCE INFORMATION ON LENIENCY IN SUBJECTIVE PERFORMANCE EVALUATION

ESSAY2. DETERMINANTS OF NARRATIVE FEEDBACK IN SUBJECTIVE PERFORMANCE EVALUATION

Essay1. The Effect of Prior Performance Information on Leniency in Subjective Performance Evaluation

I. INTRODUCTION

A long body of research in performance evaluation has investigated costs and benefits of subjective performance evaluation. Subjective performance evaluation provides useful performance information of an organization when objective measures are inadequate, incomplete, and prone to manipulation (Banker and Datar 1989; Baker et al. 1994; Bushman et al. 1996; Gibbs et al. 2004; Holmström and Milgrom 1991; Murphy 1999). Despite its importance and benefits, subjectivity has problems that supervisors make unfair and biased judgments (Baker et al. 1988; Bol 2008; Prendergast 1999). Research in psychology, organizational behavior, and accounting finds that subjectivity suffers from severe leniency and compression effects (Bol 2011; Bol et al. 2013; Moers 2005; Prendergast and Topel 1993). However, due to limited access to data, little research investigates why both effects occur and how prior performance rating affects the decisions of lenient ratings.

Leniency effect is defined as the tendency that evaluators provide performance ratings higher than those warranted by ratee performance (Saal and Landy 1977). However, it is difficult for raters to observe *warranted* performance of ratees. To estimate the warranted performance, prior empirical research employs indirect benchmarks, such as the median of subjective rating scales or the objective performance score (Bol 2011; Bretz et al. 1992; Merchant et al. 2010; Moers 2005). In subjective performance evaluations, however, all ratees perform at the different level according to their resources and their abilities (Murphy and Balzer 1989), which is inconsistent with the assumption of benchmark of the median of rating scale. Also, the objective performance score may not be an appropriate proxy for the subjective performance dimension because one of the most important roles of subjective performance measure is to modify or supplement the objective performance score (Merchant et al. 2010). Objective and subjective performance measures are supposed to represent different perspectives on performance.

Meanwhile, prior success (failure) directionally affects current expectation of success (failure) (Feather 1966). It is difficult for raters to determine whether the ratees' current performance is clearly discrepant from previous performance score (Smither et al. 1988). When the behavior of the ratees is consistent with the raters' expectations, it is automatically stored (Feldman 1981), and revealed in forms of serially correlated evaluation score. Therefore raters and ratees will create their expected score of a specific performance measure based on their prior performance level. The *prior* performance score of a specific measure is perceived as more relevant benchmark to its current performance (e.g., Thorsteinson et al. 2008; Woods 2012) than aggregated level of overall objective performance measures when both prior performance score of the same specific measure and overall objective performance level are available. Therefore, to estimate leniency, prior performance level can be a better benchmark not only for the raters but also for ratees. However, most of prior studies rely on single-year (Moers 2005), or two-year samples (Bol 2011; Merchant et al. 2010), making it almost impossible to use the prior performance information as the benchmark. I develop the proxy for leniency at the performance measure level based on prior performance level by collecting and utilizing performance evaluation data of multiple years for multiple SOEs in Korea.

In addition to the leniency measurement issue, if the leniency is expected in most subjective evaluations, the following questions will be interesting: Is leniency a one-time event or can it persist over time? If the prior subjective performance score is high, then is current performance score evaluated leniently? Once prior period subjective evaluation is lenient, then does the next period evaluation become harsh? All of the above questions cannot be explored explicitly without the time-series data. In this paper, I examine whether leniency persists over time and how prior performance information affects the incentives of raters. In doing so, I aim to provide a deeper understanding about the leniency.

When raters bias their rating toward leniency, they may estimate the expected utility of lenient rating against the disutility of the accurate rating (Murphy and Cleveland 1995, p.243). Prior studies show that raters have more incentives to rate inaccurately than accurately when they are neither penalized for biased rating nor rewarded for accurate rating (Spence and Keeping 2011), and that the incentives and characteristics of raters do not change easily. Thus, I expect that the leniency persists over time unless there is a significant change in reward (or penalty) structure for rating accuracy.

Supervisors have their own incentives to inflate ratings for low or harshly evaluated performers in the prior period so that they avoid conflicts with subordinates (Levy and Williams 2004; Mitchell and O'Reilly 1983). Also, the raters are likely to show larger leniency to poor performers than to high performers because motivation effect is greater for poor performers (Bol 2011; Kane 1994; Woods 2012). If I assume the concave utility functions of the ratees as performance score rises, the sensitivity of utility to lenient or harsh rating will be higher for the low prior performers than high prior performers. Therefore I can predict that raters show larger amount of leniency for low prior performers (or harshly evaluated performers) and smaller amount of leniency for high prior performers (or leniently evaluated performers) because raters want to reduce overall confrontation costs by maximizing total utility of ratees without violating the grading distribution in the evaluation guideline. My findings confirm the asymmetry of lenient ratings depending on the level of prior performance and the degree of prior leniency. The findings lead to additional insights that asymmetric lenient rating for low performers over time can increase the average subjective performance scores and decrease the score gap between superior performer and poor performer, resulting in both leniency and compression effects. This finding shows that the asymmetry of leniency bias is one of the contributing factors to compression bias for subjective measures. In other words, compression bias is a result of asymmetric leniency bias over time.

My study contributes to the literature in the following ways. First, I redefine the proxy for leniency based on prior performance level while prior studies have used proxies based on aggregate objective performance level or median of subjective scales. The multiple year and multiple SOE performance evaluation data enable meto develop my new definition of leniency. I construct an alternative proxy for the degree of leniency in micro-level. The proxy for leniency is based on *modified* previous performance score and reflects the expected performance level of both raters and ratees in multiple period setting. Although contract theory presents that performance is a function of both ratees' effort and noise (bias), the leniency proxies in prior studies have not explicitly discriminated between ratees' effort part and rating bias part. This paper fills this void by excluding the ratee effort portion from my leniency proxy and therefore focusing on systematic bias – leniency effect itself.

Second, this study contributes to the literature by examining the time-series pattern of leniency from the perspective of prior performance information. For example, I analyze the time series pattern of leniency responding to the necessity of time-series data study (Bol 2011; Moers 2005), and find that leniency effect persists over time. These findings increase the pervasiveness of leniency effects over time across firms. Furthermore, I extend previous research by analyzing the effects of prior performance information on the degree of raters' leniency. My findings show asymmetric leniency depending on the prior performance level and the prior leniency tendency. This indicates that the supervisors manage their dual roles as evaluators and motivators and adjust the ratings differently conditional on the ratees' past performance level. In short, my study shows that the consideration of the prior performance information is crucial in subjective evaluation, and suggests that the incentives of raters might differ with respect to their performance information set such as prior performance level and prior leniency level.

The remainder of the paper is organized as follows. I review the theoretical background and develop my hypotheses in Section II. In Section III, I provide institutional background on performance evaluation systems of SOEs in Korea. In

Section IV, I describe research design including sample selection and empirical measures. My primary empirical results are provided in Section V. Section VI concludes.

II. THEORY AND HYPOTHESIS DEVELOPMENT

2.1. Background

If the adoption of objective performance measures in incentive contracts is effective by itself in incorporating all efforts of ratees properly, any other performance signal should not be additionally valuable for contracting purposes (Holmström 1979). However, objective performance measures are often inadequate, incomplete, and prone to manipulation (Murphy 1999). Thus, to compensate for the weakness of the objective performance measure, subjective performance measure is usually combined, because it provides a more aligned incentive and reduce uncertainty to ratees by filtering out uncontrollable factors mixed in objective measures (Baker et al. 1994; Bushman et al. 1996).[1] Despite its benefit, subjective performance evaluation also has a number of problems. One example is the lenient rating. Prior studies in various disciplines such as psychology, organizational behavior, and accounting find that subjective performance evaluation suffers from severe leniency bias, the tendency of evaluators to provide performance ratings higher than those warranted by ratee performance (Bol 2011; Bretz et al. 1992; Moers 2005; Saal and Landy 1977).[2]

[1] Bol (2008) shows three different types of subjectivity: (1) using subjective performance measure, (2) allowing for flexible weighting, and (3) allowing for ex post discretion in bonus. The setting of this paper naturally controls two latter ones because the performance evaluation system prohibits supervisors from arbitrarily adjusting weight and bonus.

[2] Bretz et al. (1992) document that 60 to 70 percent of an organization's workforce are rated in the top two performance levels compared to the median of scale, the benchmark of the warranted performance level.

2.2. Hypothesis

2.2.1. Persistence of Lenient Rating

Performance evaluation literature has been premised on the assumption that rating biases come from supervisors' cognitive limitations (Murphy et al. 1982). Supervisors tend to focus more on a limited set of performance dimensions such as common measures and financial outcomes and less on unique measures and financial inputs (Ittner et al. 2003; Lipe and Salterio 2000). Specifically, research on assimilation effects and anchoring effects has shown that prior performance evaluation serves as an anchor and affects current performance rating (Huber 1989; Smither et al. 1988; Woods 2012). These studies provide evidence that cognitive limitations can deteriorate the full benefits of using subjectivity in contracting, on the assumption that the goal of raters is to rate their ratees accurately. However, recent studies argued that accuracy may not be the sole goal of raters and explain that other rater incentives cause leniency bias.

Supervisors inflate ratings of subordinates to serve their self-interests (Banks and Murphy 1985; Fried and Tiegs 1995; Longenecker et al. 1987; Spence and Keeping 2011). The probability of lenient rating is determined by the difference between raters' expected utility of biased rating and that of accurate evaluation (Murphy and Cleveland 1995). If supervisors are not rewarded for accurate ratings or punished for biases, supervisors may have insufficient motivation to invest time in gathering information (Bol 2011). Also, supervisors prefer to maintain a good relationship with their subordinates, or to minimize confrontation costs (Harris 1994; Murphy and Cleveland

Bol (2011) and Moers (2005) find that performance ratings on subjective measure are higher than on objective performance ratings, the benchmark they use.

1995), and tend to favor specific employees from political considerations (Prendergast and Topel 1993). In particular, when the rewards of supervisors are tied to the performance of subordinates, rating inflation occurs more frequently (Ilgen et al. 1981; Prendergast and Topel 1996). Golman and Bhatia (2012) also present asymmetric utility of managers as a source of leniency, suggesting that supervisors feel worse about unfavorable errors than about favorable errors.

While I expect the leniency in most subjective evaluations, only a few studies examine the time series tendency of leniency. Conducting three experiments with two-period settings, Kane et al. (1995) find that leniency tendency exists in both periods and their correlations are significant across all three experiments. Dalla Via et al. (2011) empirically examine the difference between the ratings assigned to development purposes and those for administrative purposes by using the multi-period setting. Although both studies utilize the multi-period sample, they do not explicitly analyze the pattern or the extent of leniency over time. In this study, I try to directly analyze the extent of leniency variations during the evaluation period.

The incentives of raters do not change easily unless the incentives of accurate evaluation exceed those of lenient evaluation. Without serious penalty for inaccurate rating or reward for accurate rating, the leniency effects become habitual (Murphy and Cleveland 1995). Therefore, unless there is a systematic change in evaluation guidelines against leniency overturning the utility structure of raters and ratees, I may expect the leniency persist over time. I then set up the hypothesis as follows.

H1: The leniency tendency persists over time.

2.2.2. Asymmetric Lenient Rating

Supervisors' rating incentives theory features more (less) leniency for low (high) performers and harshly (leniently) evaluated performers. In other words, asymmetric lenient rating is an outcome of self-interested rating errors because it avoids conflicts with workers (Levy and Williams 2004). If I assume a typical concave utility function for a risk-averse subordinate, marginal utility will decline as performance score rises with respect to previous performance level. Then, given the limited authority to adjust ratings and suggested rating guidelines, supervisors are more likely to inflate scores of low performers than those of high performers. Intuitively, the marginal utility of one level change in performance score is greater for low performers than high performers. Accordingly, raters are likely to inflate ratings of low performers more because the sensitivity of utility to lenient or harsh rating is greater for low past performers given the constraints of score distribution. Raters can maximize the total utility of ratees reducing the total potential confrontation costs with ratees. Also, supervisors prefer to avoid confrontation with poor performers (Mitchell and O'Reilly 1983), who tend to be less acceptable to accurate rating (Murphy and Cleveland 1995). Thus, supervisors are more likely to evaluate low performers more leniently and extend good relationships with ratees.[3] Second, from the behavioral perspective, supervisors may intentionally

[3] High past performers expect higher probability of high score in current period (Feather 1966). These expectations form and sustain the reference point level which is higher than they deserve (Abeler et al. 2011). Given the same level of harsh rating, high past performers are more likely to feel unfair than low past performers, and hence results in high confrontation cost to raters. Thus, it is possible that supervisors might also give lenient rating to high past performers. However, all ratees tend to over-estimate their abilities (Arkin et al. 1980), and hence perceive non-inflated rating as unfair regardless of their prior performance level. With a view to the theory of reference-dependent preferences, I expect that there is no difference of leniency between low past performers and high past performers.

encourage poor performers by inflating ratings than deserved, in order to influence ratees' future behavior (Bol 2011; Kane 1994; Woods 2012). Motivating low performers is considered an essential role of performance evaluation (Murphy and Cleveland 1995). Telling a poor worker that he or she is performing poorly will demotivate and discourage him or her (Longenecker et al. 1987). Therefore, targeted inaccuracy is realized as the pattern of asymmetric leniency where supervisors evaluate low performers more leniently than high performers.

If I view leniency as a result of unconscious inferential processes, however, raters' asymmetrical perception about prior score can drive asymmetric leniency. Both high performance score and low performance score produce anchoring effects when people are influenced by previous performance information that serves as an anchor for judgments (Chapman and Johnson 1999; Tversky and Kahneman 1974).[4] But the anchoring effect is larger for high performance scores than low performance scores because raters perceive low score as an extreme anchor (Thorsteinson et al. 2008). That is, the plausibility of an anchor determines the magnitude of anchoring effects (Blankenship et al. 2008; Wegener et al. 2001). Consequently, high performance score in the previous period is more likely to persist, while low performance score is more likely to be reverted to the higher score. These unintended cognitive limitations may lead to asymmetric leniency effects.

With respect to leniency in the previous period, when ratees fail to receive expected

[4] According to Tversky and Kahneman (1974), the anchoring effect is the disproportionate influence on judgment which is biased toward the initially presented value. They suggest that people make insufficient adjustments to yield a final estimation based on the anchor. In this paper, I limit the concept of the anchoring effect to an unconscious judgment, even the anchoring effect can have many different perspectives of conscious judgment.

rating, they perceive that the raters are responsible for their receiving a harsh evaluation. If subordinates believe that rewards or performance score have not been distributed equitably, the supervisor might have a strong incentive to restore perceptions of equity (Murphy and Cleveland 1995). A possible solution to restore them is to give lenient rating in the current period to the harshly evaluated subordinates in the prior period, hence raters can reduce confrontation costs of harshly evaluated ratees in the prior period.

In sum, both theories of intended and unintended rating inflation predict that raters are likely to be more lenient toward low past performers or harshly evaluated subordinates than toward high past performers or leniently evaluated ones.

> **H2a**: The leniency tendency is greater for low past performers than high past performers.
>
> **H2b**: The leniency tendency is greater for harshly evaluated performers in the previous period than leniently evaluated performers.

III. INSTITUTIONAL BACKGROUND

3.1. Overview of Performance Evaluation Systems of SOEs in Korea

The Korean government enacted the Law for Management of SOE in 2003 that requires SOE be evaluated annually by a group of auditors assigned by the government (Ministry of Strategy and Finance) and disclose the result of performance evaluation by June of the subsequent year. The government develops performance measures with which the rater and the ratee mutually agree, and distribute the guideline by the end of the year for evaluation. Then, at the end of the March, SOEs submit performance reports on the subsequent year. From April to June, raters evaluate the performance of the SOE.

The performance evaluation system employs an incentive bonus plan in which bonus size is based on the results of performance evaluation. The bonus size is limited and determined by ranking-based peer comparison via distribution analysis.[5] Thus, it is important to compare relative scores among peers. To provide ratees (SOEs) with appropriate incentives and sound evaluation systems, the Korean government has developed and actively amended the performance measurement structure for SOEs. It uses various measures under three categories: overall management, main business, and business management. Panel A in Table 1 presents an example of a performance rating of an SOE in 2005.

The third column of Panel A in Table 1 shows that each evaluation criterion classifies performance measures into objective and subjective measures. Objective measures are assessed using four quantitative, formula-based methods: actual-to-target analysis, target-range assignment analysis, trend analysis, and beta analysis.[6] In contrast to objective measures, subjective measures have a single evaluation method: grading.[7]

[5] After all the evaluations, the government transforms total performance score of SOEs into standardized Z-score. If Z-score of an SOE is above 2σ, then the SOE is graded 'S' and paid the maximum 500% bonus of monthly salary.

[6] (1) Actual-to-target analysis represents actual performance divided by target performance. Sales volume, labor expenses, and plant construction progress are examples of measures that use the actual-to-target method. (2) Target-range assignment analysis uses the ratio of actual performance minus minimum target performance to maximum target, minus minimum target performance. Capital productivity, capacity utilization, and customer satisfaction index use this evaluation method. (3) Trend analysis is a regression analysis that computes standard performance coefficients using past actual performance (e.g., the prior 15 years' figures). Actual performance is evaluated against expected performance via the standard performance coefficient. This method is, in general, used to assess how effectively SOEs manage their inventory, cost of capital, cost of goods sold, and administration expenses. (4) Beta analysis is similar to trend analysis in using past data to obtain a benchmark to assess current performance; however, it uses the beta distribution instead of the regression. Typical measures that use this method are labor productivity, economic value added (EVA), and plant power management. After conducting the quantitative assessment, each SOE is assigned into five or nine grades, in accordance with the predetermined score ranges.

[7] Subjective measures complement objective measures for activities that are difficult to quantify, but that are important to achieving firm strategic goals—for example, efficient management in strategic plans,

While objective measures are calculated by comparing actual performance based on predetermined formula for benchmark performance, assigning subjective measures to the various grade levels depends entirely on the rater's subjective evaluation with a guideline. The guideline for subjective evaluation can be summarized into two rules. First, the guideline suggests that subjective performance should be rated scored on 50 percent of performance level and 50 percent of performance improvement against the past performance level. This rule lets raters to compare current performance to previous performance for the measure. Second, it is strongly recommended that subjective performance score be distributed as following: A+/A0: 10%, B+: 10%, B0: 20%, C: 40%, D+/D0/E+/E0: 20%. This distribution guideline is applied to each SOE class and performance category.

3.2. Formation of Raters

All SOEs are categorized into 8 classes based on their business characteristics – SOC, Service, Inspection management, Culture, Industry promotion I, Industry promotion II, Education, and Pension, as shown in Panel B in Table 1.[8] Each class is audited by four groups of rating committee – one objective measure team and three subjective measure teams that grade overall management, main business, and business management. For performance evaluation of SOEs, rating committees consist mainly of academic professors, certified public accountants, and industrial experts. Raters are allocated with matrix structure in order to harmonize between class characteristics and

improvement in control systems, appropriateness in assessment and implementation of investments, development of organizational culture, appropriateness of budgeting, cooperation with community, and employee education.

[8] Panel B in Table 1 provides an example of composition of raters in 2008. The number of classification category varies six to nine as the rule changes.

measurement characteristics. For a measurement group of common measures (overall management and business management), raters consist of experts in organization management such as accounting, public administration, business, and economics. On the other hand, for a measurement group of unique measures (main business), raters are mostly professionals with experiences in the ratees' industries.

[INSERT TABLE 1 ABOUT HERE]

The government also provides various safeguards against rater bias. First, if a rater is assigned to a certain rating committee in the current year, the rater would usually move to a different committee every year to promote independence (i.e. rotation system). Second, after three years of rating service, the rater should take one year's leave of absence. Moreover, raters should sign a code of ethics in evaluation and receive a training about how to deal with interest conflicts. All these rules guarantee independence and fairness of evaluation process and reduce familiarity risk.

Raters can perform an additional on-the-spot inspection to gather supplementary information. After the inspection, score rating is normally decided unanimously by all committee members. Ratees have opportunities to express their opinions about the processes and the results of evaluation. Although the government provides the rating committee with a checklist to facilitate the subjective evaluation, it is likely to be vulnerable to leniency bias, as discussed in Section IV.

IV. RESEARCH DESIGN

4.1. Sample Selection

My data are collected from the Korean Government's annual reports on its performance evaluations for SOEs from 2005 to 2011. The reports have been released since 2004, and the first-year data are used as prior performance information in my sample.[9] I also exclude newly adopted measures due to the construction process of proxy for leniency.[10] After imposing the data requirements for the leniency computation, I obtain a final sample of 7,470 measure-SOE-year observations, representing 109 Korean SOEs.

4.2. Leniency Model

I develop a cross-sectional model for leniency and estimate Equations (1) and (2) using OLS to test if the past performance level asymmetrically affects the raters' likelihood of lenient rating:

$$LEN_{ijt} = \alpha_0 + \alpha_1 SUB\_Attain_{ij,t-1}\ (or\ LEN_{ij,t-1}) + \alpha_2 UNIQUE_{ijt} + \alpha_3\, WEIGHT_{ijt} + \alpha_4\, EXPERIENCE_{jt} + \alpha_5\, WOMAN_{jt} + \alpha_6\, SIZE_{jt} + Fixed\ Effects + \varepsilon_t. \qquad (1)$$

$$LEN_{ijt} = \alpha_0 + \alpha_1 High\_SUB_{ij,t-1}(or\ High\_LEN_{ij,t-1}) + \alpha_2\, Low\_SUB_{ij,t-1}\ (or\ Low\_LEN_{ij,t-1}) + \alpha_3\, UNIQUE_{ijt} + \alpha_4\, WEIGHT_{ijt} + \alpha_5\, EXPERIENCE_{jt} + \alpha_6\, WOMAN_{jt} + \alpha_7\, SIZE_{jt} + Fixed\ Effects + \varepsilon_t. \qquad (2)$$

where *i*, *j*, and *t* indicate performance measure *i*, firm *j*, and year *t*, respectively.

All regressions in this paper are estimated with Huber-White robust standard errors clustered by performance measure level. The standard errors are robust to both serial

[9] The Korean government enacted the Law for Management of SOEs in 1984. External raters have evaluated the performance of 13 SOEs since 1985. However, this data of performance results makes it difficult to compare inter-temporal rating tendency because (1) before and after year 2004, there is a big difference in rating systems that can mislead the effects of leniency over time, (2) in 1998, the minimum base score at 75% was eliminated, and hence there is a large difference in score rates between before-1998 and after-1998 (Ahn et al. 2010).

[10] Although I do not directly use the sample of new measure in our analysis, I utilize the distribution of new measures when I construct the alternative leniency proxies.

correlation and heteroskedasticity (Rogers 1993). I build upon related literature on both leniency and ratees' effort allocation, in order to construct the measures used in my empirical tests. First, I define my proxy for leniency and describe the factors that prior research suggests could affect leniency effects at firm-level, rater-level, and individual measure-level.

4.2.1. Leniency Proxy

Leniency effect is the tendency of evaluators to provide performance ratings higher than those warranted by ratee performance (Saal and Landy 1977). In this study, I assume that the warranted performance of subjective performance reflects the expected performance score of the raters and ratees at performance measure level. Ratees who receive positive (negative) prior performance information will have the expectation of favorable (unfavorable) ratings than the ratees who receive no prior performance information (Salvemini et al. 1993). Therefore, the expected performance score is mainly based on the previous performance level (Feather 1966; Feldman 1981; Smither et al. 1988).

Contract theory assumes that performance equals effort plus noise: $q = a + e$, where q, a, and e refers to performance, ratee's effort level, and noise, respectively. The performance q can be transformed into performance score rate (SUB_Attain_t) minus base performance, or 'benchmark' ($BENCHMARK_t$). Meanwhile, noise term e can be decomposed into two – systematic noise such as leniency bias (LEN_t) and random noise ε which is assumed to be normally distributed with zero mean and variance σ^2. Moreover, change in objective performance (ΔOBJ_Attain_t) can be a proxy for effort level a (Ahn et al. 2010; Bol 2011). In sum, I calculate the leniency proxy at the individual measure level as in Equation (3).

$$LEN_t = SUB\_Attain_t - BENCHMARK_t - \Delta OBJ\_Attain_t \text{.} \quad (3)$$

Here, the variable $BENCHMARK_t$ is supposed to represent an warranted performance level although it cannot be directly measured. Prior studies use median of subjective scale (i.e. performance score "C" in my setting) or overall objective performance score as a benchmark (e.g. Bol 2011; Bretz et al. 1992; Moers 2005). First, the benchmark of the median scale is based on the assumption that true mean level of performance corresponds to the scale midpoint. This assumption implies that all ratees perform at the same level regardless of their resources and their abilities (Murphy and Balzer 1989). However, this assumption seems illogical because (1) it is entirely possible that ratee A is better performer than ratee B depending on their efforts and abilities, (2) subjective performance evaluation is normally designed to produce a skewed distribution of performance on purpose (O'Boyle and Aguinis 2012).[11] Second, the leniency measure in Bol (2011) is defined as the difference between objective performance score and subjective performance score, which depends on the assumptions that the variance in the objective performance score is similar to the true performance variance and that performance standards are similar for the objective and subjective performance. However, this assumption is inconsistent with the purpose of the

[11] Our sample, on average, shows attainment ratio of 0.109 above the median value (grade C). This descriptive statistic implies that our sample shows leniency pattern and negatively skewed distribution (skewness = -0.244). Also, the forced distribution rule admits the skewed distribution. Therefore, our benchmark for leniency based on prior performance already incorporates skewness (Bretz et al. 1992; Kane et al. 1995; Ng et al. 2011). Our leniency proxy (*LEN*) captures the degree of leniency above guaranteed skewness which implies our leniency proxy is derived based on the assumption that performance score does not follow normal distribution.

introduction of subjective performance measure whose role is to mitigate incentive distortions. While objective performance measure is often incomplete, imperfect, and prone to manipulation (Banker and Datar 1989; Bushman et al. 1996; Holmstrom and Milgrom 1991), the introduction of subjective performance measure can improve incentive contracting because it allows value-enhancing efforts which is not captured by objective performance measure (Bol 2008). Therefore, it is not appropriate to use objective performance level as a proxy for warranted performance level of a subjective performance measure. Moreover, I estimate leniency at an individual subjective performance measure level. Overall objective score cannot be used as overall benchmark for every subjective performance measures in calculating leniency of individual subjective performance measures.

In performance evaluations in a time series setting as in my study, prior performance level will play an important role. Raters may refer to the prior performance level of a specific performance measure in forming an opinion and providing a final evaluation point. Appendix A shows empirical results that current subjective performance score is significantly affected by prior subjective performance score, not by objective performance score at both performance measure level and SOE level tests.[12] Evaluation guideline also suggests that raters should compare the current performance with the last year performance in evaluation. In other words, main reference point or benchmark for evaluation of current performance of a specific subjective performance

[12] I also compare the relevance effect of prior performance level to the current performance score with that of other benchmarks, and find that the prior performance level is more relevant to the current performance score than other benchmarks such as total prior performance and prior objective performance level in Appendix A.

measure should be prior performance level reflected in a performance score. Time-series data allows meto identify prior performance level as a benchmark.[13]

Although the prior performance score can be an improved benchmark, it also has a shortcoming because this benchmark implies that the average current rating score is just as the same as the prior performance score. If I regard the absolute score of prior performance level as the benchmark without any adjustment, the expectation level of raters and ratees can deviate from prior performance level. For example, very poor prior performers with 'zero' score are assumed to have expectation level of prior performance score (i.e. zero score), again. However, they usually expect the "mean reversion" tendency of performance score. In other words, they expect that their zero past performance score is more likely to be reverted to more higher score level than zero. If they are evaluated as zero score in the current period again, they perceive the rating is very severe. I attempt to develop a benchmark performance measure in my study, exploiting the time-series performance score data.

In my setting, "mean reversion" is *partly* incorporated by requesting the combination of 50 percent midpoint scale and 50 percent prior performance score which is based on the performance guideline of SOE. For example, the benchmark for the poor prior performers with zero score can be the sum of a half of midpoint (0.5) and a half of

[13] From the ratees point of view, they are also concerned with current performance score compared with the last year performance. Especially in our performance evaluation setting, SOEs assign the responsibility to their employees for individual performance measure. Employees assigned a specific responsibility for improving that performance score will analyze the prior performance evaluation report by the raters, remedy the shortcomings mentioned in the report, develop some positive actions to improve the current situations. All these effort are directed to increase the performance score. Those employees in charge of a specific performance measure want to influence the performance score by actions mentioned above but also by generating the *polished* performance report regarding that performance dimension. In all these efforts by raters and ratees, prior performance score will provide the most important benchmark.

prior score (0.0). Similarly, the benchmark for the superior prior performer with 1.00 score can be 0.75 (= 1.00 * 50% + 0.50 * 50%). These combination reflect distribution of top and bottom scores better.

In summary, prior subjective performance score in its raw form (SUB_Attain_{t-1}) or adjusted form ($BENCHMARK_t$), not the aggregate objective score (OBJ_Attain_t), is the critical benchmark in analyzing the persistence of leniency over time and the effect of prior performance information on leniency. In SOE setting, raters are recommended to follow the well-documented rating guideline that (1) raters should give 50 percent weight on absolute performance level and 50 percent weight on performance improvement from the prior year performance level, and (2) final performance rating is strongly required to follow a forced distribution (A+/A0: 10%, B+: 10%, B0: 20%, C: 40%, D+/D0/E+/E0: 20%). Therefore, I apply these two conditions in the guideline to estimate $BENCHMARK_t$. The example of estimation process of $BENCHMARK_t$ and LEN_t is shown in Appendix B.

This leniency proxy has intuitive appeal as it can observe biases when the actual rating distribution is deviated from the assumed distribution for the job based on the prior performance level. I calculate the leniency proxy based on the assumptions and findings that subjective performance measure tends to be independent from objective performance measure (Merchant et al. 2010), and that adjusted prior performance score, on average, stands for the assumed distribution of job at the performance measure level.[14] Also, low (high) past performance level leads to higher (lower) level of *BENCHMARK* than its original performance level. Therefore, the ratees' expectation based on "mean reversion" can be also incorporated into the measure in forms of *BENCHMARK* and effort level (ΔOBJ_Attain_t). Moreover, my leniency proxy can be generated from an individual performance measure, making possible the analysis of the incentives of both raters and ratees at the performance measure level.

4.2.2. Factors Influencing the Leniency

Prior performance level (SUB_Attain_{t-1}) is an important factor not only in measuring leniency but also in influencing leniency. The test variables of interest are the past performance level – SUB_Attain_{t-1}, Low_SUB_{t-1} and $High\_SUB_{t-1}$, and the past leniency tendency – LEN_{t-1}, Low_LEN_{t-1} and $High\_LEN_{t-1}$. The independent variable, Low_SUB_{t-1} ($High\_SUB_{t-1}$) is an indicator variable that equals to one if the score rate belongs to the lowest (highest) quartile of *SUB_Attain* in the previous year, and zero otherwise. Low_LEN_{t-1} ($High\_LEN_{t-1}$) is an indicator variable that equals to one if the score rate belongs to the lowest (highest) quartile of *LEN* in the previous year, and zero otherwise. The model also includes a comprehensive set of control variables. *UNIQUE* is an indicator variable that equals one if a performance measure is unique, and zero if it is common. Unique measures are cognitively more difficult to evaluate than common measures (Lipe and Salterio 2000). In addition, some dimensions of performance in

[14] As stated in Section III, there are separate rating committee – one objective measure team and three subjective measure teams that grade overall management, main business, and business management. This setting conceptually prevents from spillover bias which means raters' subjective evaluations are directionally influenced by an accompanying objective performance score (Bol and Smith 2011). Also, I empirically find that spillover bias is not presented in our setting. Rather, current subjective score is mainly determined by prior subjective score as shown in Appendix A.

public organizations are often hard to measure (Burgress and Ratto 2003). Therefore, raters are more likely to inflate ratings for unique measures that may incur high information gathering costs (Bol 2011). However, there is an opposite prediction regarding the effect of unique measure. Because of high comparability of common measures among ratees, raters are more pressured to inflate those ratings (Fried et al. 1999). Hence, I predict no signal on this variable. I use the weight assigned to a measure (*WEIGHT*) to proxy for the relative importance among performance measures. Raters are more pressured to inflate ratings of highly weighted measures due to higher confrontation costs. I expect measurement weight to be positively associated with the tendency toward leniency. *EXPERIENCE* measures the experience of raters with the average tenure of raters at a rating committee. As tenure of raters increases, favoritism may increase, and hence raters are more likely to inflate rating (Prendergast and Topel 1996). However, rotation rule and sabbatical year system in my sample prevent raters from exercising favoritism. Research in psychology also shows that greater degrees of expertise and experience raters make evaluation of ratees more reliable and less lenient (Brown 1968; Schneier 1977). Therefore, I expect a negative relation between experience of raters and the lenient ratings. *WOMAN* measures the gender effect of raters as the number of female raters divided by total number of raters within a class. In a number of studies in performance evaluation, the gender effects of rater have been considered possible sources of variance in ratings (Landy and Farr 1980). Although some laboratory studies find that female raters are more lenient than male raters (Bartol and Butterfield 1976), others show that gender does not typically affect evaluations

(Peters et al 1984; Pulakos and Wexley 1983). Therefore, I do not have a signed prediction on the effect of woman raters on the leniency. *SIZE* is included because most previous studies support that large firms have better performance scores because they have sufficient amount of resources and capabilities, more knowledge or experiences, and well-established system (Lee 2009; Nooteboom 1993). The higher scores may partially come from leniency effect (Ahn et al. 2011). Large firms with abundant resources prepare refined reports and present them in a polished manner. Also, raters may face larger confrontation costs from larger firms. Thus, large firms are likely to get higher score ratings than small firms do. On the other hand, there is a great deal of pressure on raters to evaluate large firms accurately. SOEs with more employees receive more public attention and media coverage (Du et al. 2012), and hence, raters and the government are more cautious when they evaluate and inspect the performance of large firms.. These conflicting findings of prior studies preclude mefrom making a signed prediction on the effect of firm size on leniency.

V. EMPIRICAL RESULTS

5.1. Descriptive Statistics

Figure 1, Panel A shows the histogram of subjective performance score rate (*SUB_Attain*). The mean (median) of subjective performance score rate is 0.664 (0.650) and the standard deviation is 0.131. Figure 1, Panel B presents that the mean (median) of objective performance score rate is 0.880 (1.000) and the standard deviation is 0.205. In my sample data, the mean and median rating achievement ratio of the subjective measures are lower than those of objective measures. This is inconsistent with the conventional leniency bias that has been documented in the prior literature that measures leniency by comparing subjective performance score to objective performance score (Bol 2011; Moers 2005). Participative or formula-based target setting process for objective performance measure may allow the ratees to set easier targets and to reduce efforts after the targets are met, especially when raters are not residual claimants as in the setting of this paper (e.g. Ahn and Choi 2010; Merchant and Manzoni 1989).

[INSERT FIGURE 1 ABOUT HERE]

I present summary statistics for the leniency, subjective performance score rate, and characteristics of measurement, rater, and firms in Table 2. The average (median) leniency (*LEN*) is 0.025 (0.024) and its standard deviation is 0.143. More importantly, the positive value of *LEN* persists over the sample period, as shown in Figure 2, Panel A and all the values of *LEN* are statistically different from zero and positive except for year 2005. These descriptive statistics and t-test results present that leniency persists over time in my sample period, which supports H1.[15]

The asymmetric leniency may also induce compression effects, the tendency of raters to provide performance ratings that fail to distinguish between ratees (Ahn et al. 2010; Bol 2011). If high (low) past performers are continuously rated lower (higher) in subsequent periods, then the distribution of rating will be concentrated around the median, and compression effects increase over time. As shown in Figure 2, Panel B, the

[15] The leniency cannot infinitely persist over time because of the score constraint. Therefore, the continuous leniency might be controlled in several ways by prohibiting leniency (mandatory forced distribution) and changing performance evaluation system and measures. Ahn and Kim (2013) find one of the possibilities that leniently rated subjective performance measures are likely to be dropped. Other possibilities are beyond scope of this paper, and remain as a subject of future study.

standard deviations of both *LEN* and *SUB_Attain* monotonically decrease over time, representing the time-series compression effects.[16]

Subjective performance score rate (*SUB_Attain*) monotonically increases over time, with the mean of Δ*SUB_Attain* at 2.1 percent. I conjecture that the positive *LEN* represents the time trend of subjective performance score rate (*SUB_Attain*). The objective performance score rate (*OBJ_Attain*) shows relatively stable pattern over time. The tenure of raters is, on average, about two years. The average fraction of woman raters is 7.5 percent, indicating that most raters consist of male raters. Table 3 presents the Pearson correlation matrix. In general, the control variables are not highly correlated with each other.

[INSERT FIGURE 2 ABOUT HERE]

[INSERT TABLE 2 ABOUT HERE]

[INSERT TABLE 3 ABOUT HERE]

5.2. Impact of Prior Performance Information on Leniency

Columns (1) to (6) of Table 4 present the estimation results of Equations (1) and (2), where the leniency is the dependent variable. I replace prior subjective performance score rate (SUB_Attain_{t-1}) with indicator variables of the highest and the lowest quartile of prior subjective performance score rate ($High\_SUB_{t-1}$ and Low_SUB_{t-1}) in column (2). I find that measures with higher (lower) past performance score rate tend to exhibit a lower (higher) leniency, consistent with H2a. The negative relation between the

[16] I find the significant decreasing time trend by estimating the regression. When I regress annual and industrial average standard deviations of *LEN* (*SUB_Attain*) on time trend, the coefficient is -0.005 (-0.007) at p-value < 1%.

previous performance level and leniency supports the notion that raters will positively inflate ratings of low-performing ratees. Also, in column (4) and (5) of Table 4, I find that measures with leniently (harshly) rated performance score rate tend to show a lower (higher) leniency, consistent with H2b. These results show the *reverting* tendency of leniency because raters have incentives to restore perceptions of equity.[17]

With respect to control variables, I find that leniency is significantly higher in unique measures than in common measures. The superior requires specific knowledge about an individual subordinate's characteristics to correctly evaluate the performance of the unique measure. The high information gathering costs let raters promote leniency (Bol 2011). I also find that measures of high *WEIGHT* are more likely to be rated leniently. Moreover, firms with a more experienced raters receive a lower leniency in column (1) and (2), stressing the importance of raters' characteristics in interpreting the leniency effects. Firm size is negatively related to leniency in column (4) and (5). These results indicate that raters are misled by firm size or not sophisticated enough to evaluate the performance of large firms. However, the gender effect on leniency is statistically insignificant.

The results in Columns (1), (2), (4) and (5) in Table 4 do not address the endogenous nature of the leniency. If some of the *unobserved* measurement characteristics of leniency are also determinants of the previous performance, then the previous performance score rate may spuriously affect leniency. To address the effect of endogenously determined leniency, I estimate Equation (2) using a performance

[17] I can also use a median of subjective measure scale (performance score "C") as a benchmark performance. I empirically test and find the similar results to the main findings of this paper.

measure fixed-effects model to control for time-invariant unobserved heterogeneity. As shown in column (3) and (6) of Table 4, the coefficients on *High_SUB*$_{t-1}$, *Low_SUB*$_{t-1}$, *High_LEN*$_{t-1}$, and *Low_LEN*$_{t-1}$ continue to be significant at the 1 percent level, suggesting that my results are largely robust to corrections for the endogenous nature of the leniency. In sum, I interpret these results as confirming my previous findings of a negative relation between the prior performance score rate, the prior leniency tendency and the current tendency toward leniency.

[INSERT TABLE 4 ABOUT HERE]

5.3. Additional Analysis

5.3.1. Alternative Proxies for Leniency

I check the robustness of my analysis on the asymmetric leniency by using the alternative leniency proxies. The previous results of estimating the relation between leniency and prior performance are based on the two government guidelines for SOE evaluation—(1) to use a forced distribution, and (2) to incorporate 50 percent of performance level and 50 percent of year-on-year performance improvement. To test the robustness of asymmetric leniency effects, I relax the two assumptions one by one.

First, raters can choose not to follow the forced distribution guideline although it is strongly recommended.[18] I employ "the distribution of new measure", the proxy for the actual rating tendency of raters.[19] I utilize the distribution of new measures on annual base and apply the distribution to calculate *BENCHMARK2*. Then I calculate leniency

[18] Merchant et al. (2010) find no evidence of a leniency effect in the subjective performance ratings, when forced distribution is mandated.

[19] Our database consists of existing performance measures and newly adopted performance measures. In the previous analysis, I only utilize existing performance measures because the consecutive two-year observations are necessary to calculate our proxy for leniency.

(*LEN2*) by using the same Equation (3). Second, I assume that change in objective performance score rate represents rater efforts. When I eliminate the change in objective performance score from Equation (3), I get the additional proxy for leniency, which is free from the effort adjustment. In other words, *LEN3* is calculated by extracting the change in objective performance score rate part (Δ*OBJ_Attain*) from the original leniency (*LEN*).[20] Third, even though the "50 percent of performance level and 50 percent of performance improvement" rule is mandatory when I calculate the *BENCHMARK*, I relax this rule for the purpose of the robustness check. Then the benchmark is 100 percent of prior performance score level. If change in objective performance score rate is also ignored, then change in subjective performance score rate itself ($\Delta SUB\_ Attain_t$) is the proxy for leniency. Finally, to check the possibility that raters think of not only one-year prior performance but also prior performance trend as the benchmark, I use the average performance score rate in the previous three years as the benchmark ($SUB\_ Attain2_t$). I incorporate these four alternative measures of leniency into the Equation (2) as shown in the Table 5. The results show the evidence consistent with my prediction that lower (higher) past performance is positively (negatively) associated with lenient ratings.[21] In an untabulated analysis, I also find the reverting pattern of leniency, consistent with H2b.

[20] I use change in objective performance score as a proxy for the subordinates' effort level regarding the subjective performance measure. However, as stated above, objective and subjective performance measures are supposed to represent different perspectives on performance. Thus, when I replace change in objective performance score (ΔOBJ_Attain_t) with change in subjective performance score at SOE level (ΔSUB_Attain_t), the results remain the same.

[21] The previous results with our original leniency proxy are more conservative than those with alternative proxies in terms of statistical significance. Benchmark with prior score only can induce much higher effect of past performance level on leniency.

[INSERT TABLE 5 ABOUT HERE]

5.3.2. Alternative Proxies for Prior Performance

My leniency proxy is based on subjective performance score in the previous period. However, the prior performance score might be also subject to leniency bias because the performance score in the prior period also reflects subordinate's effort and leniency bias as shown in Equation (3). Thus, if the prior subjective performance score is regarded as the benchmark of current subjective performance rating, my leniency proxy might capture the leniency biases of prior periods as well as the current period. Specifically, I measure leniency as difference of current to past performance, and the past performance encompasses past leniency bias. Continuing the same process backward, my leniency proxy equals the current leniency bias minus leniency bias in the previous year plus that of two years ago and so on. In my first robustness check, I restrict my analysis to SOEs in 2005 which are initially evaluated in 2004 and are free from serial correlation of leniency. In an untabulated analysis, I find the same asymmetric leniency pattern – lenient (harsh) for low (high) performers.[22]

I additionally check the possibility of another standard of judgment about prior performance. Raters normally have information of both past subjective performance scores and current objective performance scores when they conduct evaluation of subjective performance measures. In other words, the overall objective performance rates of current year are likely to be known by raters before subjective performance evaluation. Therefore, there is a possibility that raters would judge ratees' performance

[22] The coefficient of Low_SUB_t and $High\_SUB_t$ is 0.020 at p-value < 10% and -0.028 at p-value < 1%, respectively.

level by their current objective performance rate and evaluate ratees leniently or harshly according to their incentives. In an untabulated analysis, I find the same asymmetric leniency pattern – lenient (harsh) for low (high) performers – when I classify superior and poor performer by using current objective performance score.[23]

5.3.3. Same Rater Bias

Typical argument is that if a rater consecutively evaluates the same ratee for two periods, familiarity and favoritism will increase, resulting in lenient rating (Prendergast and Topel 1996). It is impossible, however, to test this possibility in my setting. The rater rotation rule is mandated in Korean SOE setting, and thus, the same rater cannot evaluate the same SOE for consecutive two years. Alternatively, I manually trace all the raters' evaluation committee, not individual level, and make a proxy for the fraction of raters who evaluate the same industry for consecutive two years. Untabulated results suggest that there is statistically significant and negative relationship between the portion of raters who rate the same industry for consecutive two years and lenient rating tendency when I add the proxy in Equation (2). In other words, similar to *EXPERIENCE* measure, raters who have experience with evaluating a SOE in the same industry are more likely to evaluate ratees more reliably and less leniently because of reduced information gathering costs (Brown 1968; Schneier 1977).

VI. CONCLUSION

[23] When I replace the past subjective performance (SUB_Attain_{t-1}) with the current objective performance score rate (OBJ_Attain_t) as a proxy for the superiority of ratee's performance, the result remains the same. The coefficient of the OBJ_Attain_t is -1.040 at p-value < 1%.

The purpose of this paper is to examine how leniency behaves over time in terms of persistence, whether lenient ratings are affected by the past performance information. Using the dataset of multiple SOEs for multiple years, I construct the proxy for leniency at the performance measure level, based on the prior performance information. I find that (1) leniency effect persists over time, (2) leniency is reverted, and (3) high (low) past performers tend to be rated less (more) leniently. Also, these findings are robust to four alternative proxies for leniency. Interestingly the persistence of leniency over time and asymmetric leniency contributes to compression bias. Leniency and compression effects do not stem from different motivations, but from the same incentive to evaluate the low performers more leniently.

The results of this study have several practical implications. I examine the incentive and reaction of raters with regard to lenient rating. My study shows that the consideration of the prior performance information is important in subjective evaluation, and suggests that the incentives of raters and ratees differ with respect to their previous performance level and their prior leniency. Also, the intrinsic motivation of public sector employees is a major premise condition of effort and performance (e.g., Prendergast 2007; Wright 2001), and is higher than that of private sector employees (Lyons et al. 2006). Given its importance in this sector, my results suggest that raters should comprehensively consider the motivational effects of leniency in subjective evaluation focusing on the prior performance information.

My findings are subject to important caveats. Most importantly, generalizability of this study is limited because this study is based only on public entities in Korea and the quality of the variables employed in this study may not be optimal. Societal norms and values in Korea may also have some influence on my results. However, my sample firms employ a wide range of performance measures (e.g., subjective versus objective, common versus unique) and adopt a performance-based bonus system that is parallel to performance evaluation system in most successful enterprises. Moreover, I find the leniency effect in spite of several anti-leniency policies in my setting – such as rotation rule, sabbatical year system about raters, and forced distribution rule about performance score. This finding suggests that the persistence of leniency and asymmetric leniency toward past performance information might be more frequently observed in other normal setting. Second, my proxy for leniency is constructed based upon the field-specific assumptions in Korean SOEs. Therefore, I should exercise caution when generalizing the results based on my leniency proxy. Third, I assume that subjective performance measure might complement the shortcomings of objective performance measure. However, the relationship between subjective and objective performance measure can be considered empirical research subject (Merchant et al. 2010). Lastly, even though I suggest that leniency persists over time, the leniency tendency cannot infinitely persist because of the score constraint. Future research can address these limitations by incorporating the age of performance measure and the timing of performance measure drop.

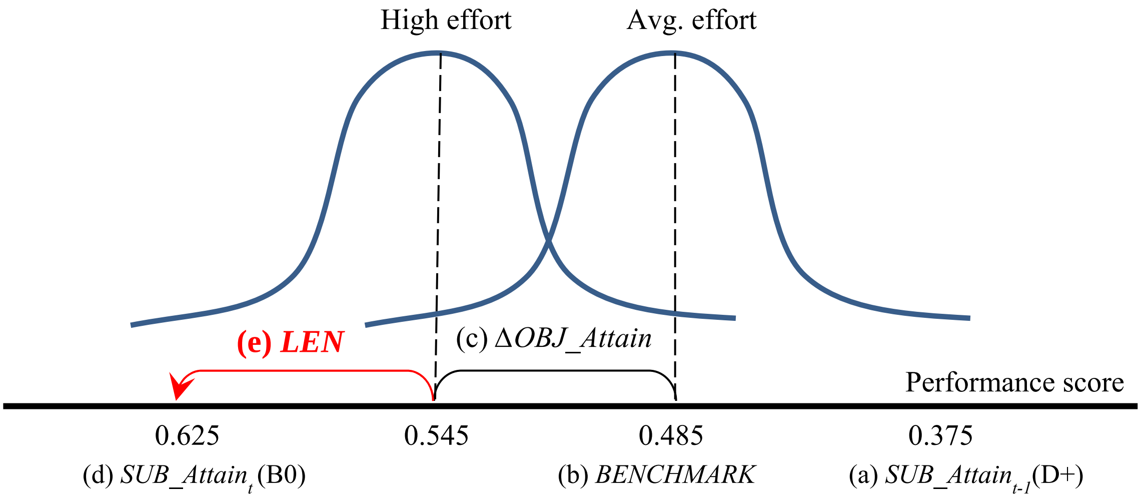
APPENDIX A
Test for Relevance of Benchmarks to Current Subjective Score

		(1)	(2)	(3)	(4)	(5)	(6)
		PM level	PM level	SOE level	SOE level	SOE level	SOE level
Independent Variables [b]	Predicted Sign	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient
		(t-value)	(t-value)	(t-value)	(t-value)	(t-value)	(t-value)
SUB_Attain_{t-1}	+	0.337***	0.311***	0.634***	0.465***	0.614***	0.558***
		(14.90)	(8.81)	(22.23)	(8.75)	(10.78)	(2.83)
OBJ_Attain_{t-1}	+	0.052***	-0.010	0.026	0.016	0.024	0.013
		(3.61)	(-0.49)	(0.89)	(0.45)	(0.80)	(0.38)
TOT_Attain_{t-1}	+					0.030	-0.152
						(0.41)	(-0.45)
Controls		$UNIQUE_t$ $WEIGHT_t$ $EXPERIENCE_t$ $WOMAN_t$ $SIZE_t$	$UNIQUE_t$ $WEIGHT_t$ $EXPERIENCE_t$ $WOMAN_t$ $SIZE_t$ $\boldsymbol{SUB\_Attain_{t-2}}$ $\boldsymbol{SUB\_Attain_{t-3}}$	$EXPERIENCE_t$ $WOMAN_t$ $SIZE_t$	$EXPERIENCE_t$ $WOMAN_t$ $SIZE_t$ $\boldsymbol{SUB\_Attain_{t-2}}$ $\boldsymbol{SUB\_Attain_{t-3}}$ $\boldsymbol{OBJ\_Attain_{t-1}}$ $\boldsymbol{OBJ\_Attain_{t-2}}$	$EXPERIENCE_t$ $WOMAN_t$ $SIZE_t$	$EXPERIENCE_t$ $WOMAN_t$ $SIZE_t$ $\boldsymbol{SUB\_Attain_{t-2}}$ $\boldsymbol{SUB\_Attain_{t-3}}$ $\boldsymbol{OBJ\_Attain_{t-1}}$ $\boldsymbol{OBJ\_Attain_{t-2}}$ $\boldsymbol{TOT\_Attain_{t-2}}$ $\boldsymbol{TOT\_Attain_{t-3}}$
Intercept		0.321***	1.140***	0.207***	0.242***	0.200***	0.228***
		(2.81)	(4.45)	(6.69)	(4.70)	(5.93)	(4.62)
Year Fixed Effects		Yes	Yes	Yes	Yes	Yes	Yes
SOE Fixed Effects		Yes	Yes				
SOE Clustering				Yes	Yes	Yes	Yes
Measure Clustering		Yes	Yes				
Number of observations		7,470	3,310	500	295	500	295
Adjusted R^2		0.443	0.378	0.689	0.596	0.688	0.593

T-statistics are reported in parentheses under each estimated coefficient. The symbols *, **, and *** correspond to 10 percent, 5 percent, and 1 percent significance levels, respectively, for two-tailed t-tests. TOT_Attain_{t-1} is defined as the prior total performance score rate of a SOE. See Table 2 for the definitions of other variables.

APPENDIX B
Approximation of Leniency Measure (LEN)

Panel A: Numerical Example of Estimation Process of *LEN*



Panel B: Example of *BENCHMARK* in 2005

Prior Grade	50% of Level	50% of Improvement	*BENCHMARK*
A+ (1.000)	40% A+ = 0.400 40% A+ = 0.400 20% A0 = 0.175 0.975	0.569	50% * 0.975 50% * 0.569 ---------------- 0.772
A0 (0.875)	40% A+ = 0.400 40% A0 = 0.350 20% B+ = 0.150 0.900	0.569	50% * 0.900 50% * 0.569 ---------------- 0.735
…	…	…	…
E+ (0.125)	40% D0 = 0.100 40% E+ = 0.050 20% E0 = 0.000 0.150	0.569	50% * 0.150 50% * 0.569 ---------------- 0.360
E0 (0.000)	40% E+ = 0.050 40% E0 = 0.000 20% E0 = 0.000 0.050	0.569	50% * 0.050 50% * 0.569 ---------------- 0.310

Raters are supposed to follow these guidelines while keeping the prior performance score in mind as the reference point. First, '50 percent of improvement' part would be applied by the forced distribution rule (A+/A0: 10%, B+: 10%, B0: 20%, C:40%, D+/D0/E+/E0: 20%). Because this part should be independently determined and not be affected by the previous performance score in accordance with the rule, "50 percent of improvement" portion for every measure has the same distribution and weighted average of each rating of about 0.569. The subjective performance rating has nine grades, from A+ to E0. This rating determines score rates: a grade of A+ translates to the score rate of 100 percent, diminishing by 12.5 percent for each lower grade, and ending at 0 percent for E0. Thus, multiplying the score level by its weight which is based on the forced distribution rule yields 0.569 [= 1.000 (A+) * 5% + 0.875 (A0) * 5% + 0.750 (B+) * 10% + 0.625 (B0) * 20% + 0.500 (C) * 40% + 0.375 (D+) * 20%). Second, the calculation of "50 percent of performance level" part starts from the prior score level and slightly adjusts the score according to "50 percent of improvement" part. In other words, the current performance level can be upgraded or degraded or unchanged from the prior score rating. As a numerical example in Panel A, let's assume that the previous subjective performance score (*SUB_Attain*$_{t-1}$) of "Organization management" measure in a SOE is D+ (0.375). If raters evaluate that performance improves with a probability of 40 percent (i.e. the sum of A+/A0: 10%, B+: 10%, B0: 20%), then performance level will be upgraded by one level, (i.e. C: 0.50). No change in performance with probability of 40 percent (C: 40%) earns grade D+ (0.375), and deteriorated performance earns grades below D+ in 50 percent of improvement, and performance level will be degraded by one level D0 (0.25) with probability of 20 percent. Hence, the "50 percent of level" part is the sum of the three cases, 0.40 [= 0.5 (C) * 40% + 0.375 (D+) * 40% + 0.25 (D-) * 20%] which is mainly subject to the prior performance level. So, the benchmark level ($BENCHMARK_t$) is 0.485, equals to the sum of 1) the weighted average 0.285 from "50 percent of improvement" part, and 2) 0.200 from "50 percent of performance level" part. Panel B presents the example of benchmark level ($BENCHMARK_t$) in 2005. Annual adjustment to the benchmark is necessary because of intermittent rule change in scaling. This produces a benchmark table at 'grade-year' level that has, on average, nine benchmarks per year.

Also, if I assume that *current* performance achievement of "Organization management" measure is B0 (0.625) and changes in objective performance achievement is 0.060, then the value of leniency (LEN_t) of the performance measure is 0.080 [= SUB_Attain_t (0.625) – $BENCHMARK_t$ (0.485) – ΔOBJ_Attain_t (0.060)].

REFERENCES

Abeler, J., A. Falk, L. Goette, and D. Huffman. 2011. Reference points and effort provision. *American Economic Review* 101: 470-492.

Ahn, T. S., and Y. S. Choi. 2010. Performance management under dynamic incentive scheme. Working paper, Seoul National University.

Ahn, T. S., I. Hwang, and M. Kim. 2010. The impact of performance measure discriminability on rate incentives. *The Accounting Review* 85(2): 389-417.

Ahn, T. S., and B. J. Kim. 2013. Why do performance measures drop? Working paper, Seoul National University.

Ahn, T. S., J. Y. Lee, and J. H. Park. 2011. The effects of firm size and measurement characteristics on performance score. *Korean Journal of Management Accounting Research* 11(2): 107-137.

Arkin, R., H. Cooper, and T. Kolditz. 1980. A statistical review of the literature concerning the self-serving attribution bias in interpersonal influence situations. *Journal of Personality* 48(4): 435-448.

Baker, G. P., R. Gibbons, and K. J. Murphy. 1994. Subjective performance measures in optimal incentive contracts. *The Quarterly Journal of Economics* 109(4): 1125-1156.

Baker, G. P., M. C. Jensen, and K. J. Murphy. 1988. Compensation and incentives: Practice vs. theory. *Journal of Finance* 43(3): 593-616.

Banker, R. D., and S. M. Datar. 1989. Sensitivity, precision, and linear aggregation of signals for performance evaluation. *Journal of Accounting Research* 27(1): 21-39.

Banks, C., and K. Murphy. 1985. Toward narrowing the research-practice gap in performance appraisal. Personnel Psychology 38(2): 335−345.

Bartol, K. M., and D. A. Butterfield. 1976. Sex effects in evaluating leaders. *Journal of Applied Psychology* 61: 446-454.

Blankenship, K. L., D. T. Wegener, R. E. Petty, B. Detweiler-Bedell, and C. L. Macy. 2008. Elaboration and consequences of anchored estimates: an attitudinal perspective on numerical anchoring. Journal of Experimental Social Psychology 44(6): 1465–1476.

Bol, J. 2008. Subjectivity in compensation contracting, *Journal of Accounting Literature* 27: 1-24.

Bol, J. 2011. The determinants and performance effects of managers' performance evaluation biases. *The Accounting Review* 86(5): 1549-1575.

Bol, J., S. Kramer, V. Maas, and S. Richtermeyer. 2013. Managers' Incentives in the Performance Evaluation Process: The Role of Information Accuracy and Bonus Transparency. AAA 2014 Management Accounting Section (MAS) Meeting Paper.

Bol, J. C., and S. D. Smith. 2011. Spillover effects in subjective performance evaluation: Bias and the asymmetric influence of controllability. *The Accounting Review* 86(4): 1213-1230.

Bretz, R. D., G. T. Milkovich, and W. Read. 1992. The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management* 18(2): 321-352.

Brown, E. M. 1968. Influence of training, method, and relationship on the halo effect. *Journal of Applied Psychology* 52: 195-199.

Burgess, S., and M. Ratto. 2003. The role of incentives in the public sector: Issues and evidence. *Oxford Review of Economic Policy* 19(2): 285-300.

Bushman, R. M., R. J. Indjejikian, and A. Smith. 1996. CEO compensation: The role of individual performance evaluation. *Journal of Accounting and Economics* 21(2): 161-193.

Chapman, G. B., and E. J. Johnson. 1999. Anchoring, activation, and the construction of values. *Organizational Behavior and Human Decision Processes* 79(2): 115-153.

Dalla Via, N., F. Hartmann, and P. Collini. 2011. Long-term incentives, managerial effort and supervisor evaluation bias. Working paper, Erasmus University Rotterdam.

Du, F., G. Tang, and S. M. Young. 2012. Influence activities and favoritism in subjective performance evaluation: Evidence from Chinese state-owned enterprises. *The Accounting Review* 87(5): 1555-1588.

Feather, N. T. 1966. Effects of prior success and failure on expectations of success and subsequent performance. *Journal of Personality and Social Psychology* 3(3): 287-298.

Feldman, D. C. 1981. The multiple socialization of organization members. *Academy of Management Review* 6(2): 309-318.

Fried, Y., S. Ariel, A. S. Levi, H. A. Ben-David, and R. B. Tiegs. 1999. Inflation of subordinates' performance ratings: main and interactive effects of rater negative affectivity, documentation of work behavior, and appraisal visibility. *Journal of Organizational Behavior* 20(4): 431-444.

Fried, Y., and R. B. Tiegs. 1995. Supervisors' role conflict and role ambiguity differential relations with performance ratings of subordinates and the moderating effect of screening ability. *Journal of Applied Psychology* 80(2): 282-291.

Gibbs, M., K. A. Merchant, W. A. Van der Stede, and M. E. Vargus. 2004. Determinants and effects of subjectivity in incentives. *The Accounting Review* 79(2): 409-436.

Golman, R. and Bhatia, S. 2012. Performance evaluation inflation and compression. *Accounting, Organizations and Society* 37: 534-543.

Harris, M. M. 1994. Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management* 20(4): 737-756.

Holmström, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10(1): 74-91.

Holmström, B., and P. Milgrom. 1991. Multi-task principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization* 7: 24-52.

Huber, V. L. 1989. Comparison of the effects of specific and general performance standards on performance appraisal decisions. *Decision Sciences* 20: 545-557.

Ilgen, D. R., T. R. Mitchell, and J. W. Frederickson. 1981. Poor performers: Supervisors' and subordinates' responses. *Organizational Behavior and Human Performance*, 27: 386-410.

Ittner, C. D., D. F. Larcker, and M. Meyer. 2003. Subjectivity and the weighting of performance measures: Evidence from a balanced scorecard. *The Accounting Review* 78(3): 725-758.

Kane, J. S. 1994. A model of volitional rating behavior. *Human Resource Management Review* 4(3): 283-310.

Kane, J. S., H. J. Bernardin, P. Villanova, and J. Peyrefitte. 1995. Stability of rater leniency: Three studies. *The Academy of Management Journal* 38(4): 1036-1051.

Landy, F. J., and J. L. Farr. 1980. Performance rating. *Psychological Bulletin* 87(1): 72-107.

Lee, J. 2009. Does Size Matter in Firm Performance? Evidence from US Public Firms. *International Journal of the Economics of Business* 16(2): 189-203.

Levy, P. E., and J. R. Williams. 2004. The social context of performance appraisal: A review and framework for the future. *Journal of Management* 30(6): 881-905.

Lipe, M. G., and S. E. Salterio. 2000. The balanced scorecard: Judgmental effects of common and unique performance measures. *The Accounting Review* 75(3): 283-298.

Longenecker, C., H. Sims, and D. Gioia. 1987. Behind the mask: The politics of employee appraisal. *Academy of Management Executive* 1: 183-193.

Lyons, S.T., Duxbury, L.E. and Higgins, C.A., 2006. A comparison of the values and commitment of private sector, public sector, and parapublic sector employees. *Public Administration Review* 66(4): 605–618.

Merchant, K. A., and J. Manzoni. 1989. The achievability of budget targets in profit centers: A field study. *The Accounting Review* 64(3): 539-558.

Merchant, K. A., C. Stringer, and P. Theivananthampillai. 2010. Relationships between objective and subjective performance ratings. Working paper, University of Southern California.

Mitchell, T. R., and C. A. O'Reilly III. 1983. Managing poor performance and productivity in organizations. *Research in Personnel and Human Resources Management* (1): 201-234.

Moers, F. 2005. Discretion and bias in performance evaluation: The impact of diversity and subjectivity. *Accounting, Organizations and Society* 30(1): 67-80.

Murphy, K. J. 1999. Executive compensation. *Handbook of Labor Economics* 3: 2485-2563.

Murphy, K. J., and W. K. Balzer. 1989. Rater Errors and Rating Accuracy. *Journal of Applied Psychology* 74: 619-624.

Murphy, K. J., W. K. Balzer, M. C. Lockhart, and E. J. Eisenman. 1985. Effects of previous performance on evaluations of present performance. *Journal of Applied Psychology* 70(1): 72-84.

Murphy, K. R. and J. N. Cleveland. 1995. *Understanding Performance Appraisal*, Thousand Oaks, CA: Sage Publications.

Murphy, K. R., M. Garcia, S. Kerkar, C. Martin, and W. K. Balzer. 1982. Relationship between observational accuracy and accuracy in evaluating performance. *Journal of Applied Psychology* 67: 320-325.

Ng, K., C. Koh, S. Ang, J. C. Kennedy, and K. Chan. 2011. Rating leniency and halo in multisource feedback ratings: Testing cultural assumptions of power distance and individualism–collectivism. *Journal of Applied Psychology* 96(5): 1033-1044.

Nooteboom, B. 1993. Firm Size Effects on Transaction Costs. *Small Business Economics* 5(4): 283-295.

O'Boyle, E., and H. Aguinis. 2012. The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology* 65(1): 79-119.

Peters, L. H., E. J. O'Connor, J. Weekley, A. Pooyan, B. Frank, and B. Erenkrantz. 1984. Sex bias and managerial evaluations: A replication and extension. *Journal of Applied Psychology* 69(2): 349-352.

Prendergast, C. 1999. The provision of incentives in firms. *Journal of Economic Literature* 37(1): 7-63.

Prendergast, C., 2007. The motivation and bias of bureaucrats. *American Economic Review* 97(1): 180–196.

Prendergast, C., and R. H. Topel. 1993. Discretion and bias in performance evaluation. *European Economic Review* 37(2-3): 355-365.

Prendergast, C., and R. H. Topel. 1996. Favoritism in organizations. *Journal of Political Economy* 104(5): 958-978.

Pulakos, E. D., and K. N. Wexley. 1983. The relationship among perceptual similarity, sex, and performance ratings in manager-subordinate dyads. *Academy of Management* 26(1): 129-139.

Rogers, W. 1993. Regression standard errors in clustered samples. *Stata Technical Bulletin* 13: 19-23.

Saal, F. E., and F. J. Landy. 1977. The mixed standard rating scale: An evaluation. *Organizational Behavior and Human* Performance 18: 19-35.

Salvemini, N. J., R. R. Reilly, and J. W. Smither. 1993. The influence of rater motivation on assimilation effects and accuracy in performance ratings. *Organizational Behavior and Human Decision Processes* 55(1): 41-60.

Schneier, C. E. 1977. Operational utility and psychometric characteristics of BES: A cognitive reinterpretation. *Journal of Applied Psychology* 62(5): 541-548.

Smither, J. W., R. R. Reilly, and R. Buda. 1988. Effect of prior performance information on ratings of present performance: Contrast versus assimilation revisited. *Journal of Applied Psychology* 73: 487-496.

Spence, J. R., and L. M. Keeping. 2011. Conscious rating distortion in performance appraisal: A review, commentary, and proposed framework for research. *Human Resource Management Review* 21: 85-95.

Strack, F., and T. Mussweiler. 1997. Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology* 73 (3): 437–446.

Thorsteinson, T. J., J. Breier, A. Atwell, C. Hamilton, and M. Privette. 2008. Anchoring effects on performance judgments. *Organizational Behavior and Human Decision Processes* 107: 29-40.

Tversky, A., and D. Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185: 1124-1131.

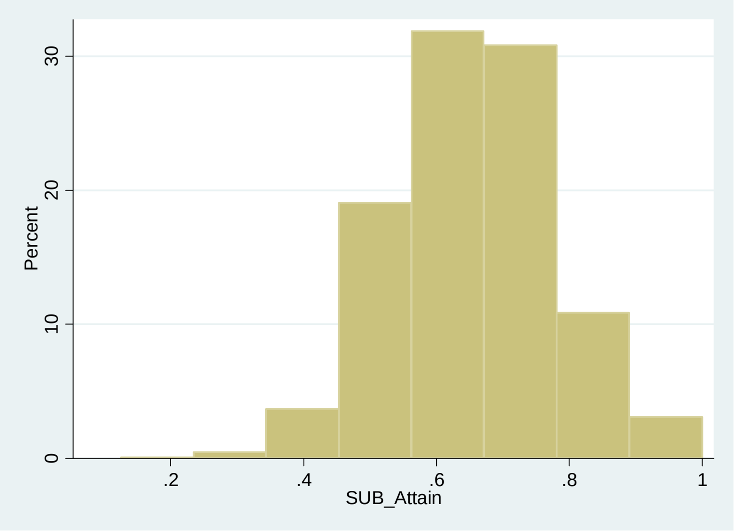
Wegener, D. T., R. E. Petty, B. T. Detweiler-Bedell, and W. B. G. Jarvis. 2001. Implications of attitude change theories for numerical anchoring: Anchor plausibility and the limits of anchor effectiveness. *Journal of Experimental Social Psychology* 37: 62-69.

Woods, A. 2012. Subjective adjustments to objective performance measures: The influence of prior performance. *Accounting, Organizations and Society* 37: 403-425.

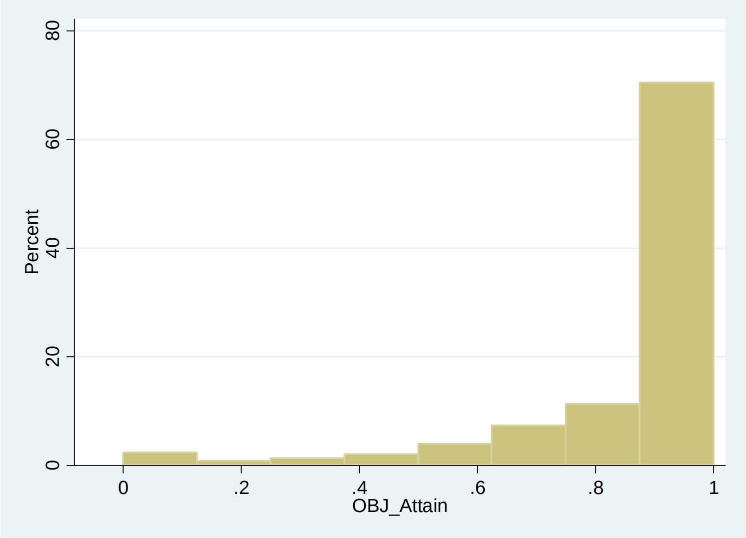
Wright, B. E., 2001. Public-sector work motivation: a review of the current literature and a revised conceptual model. *Journal of Public Administration Research and Theory* 11(4): 559–586.

FIGURE 1
Distributions of Score Rate

Panel A: Distribution of *SUB_Attain*

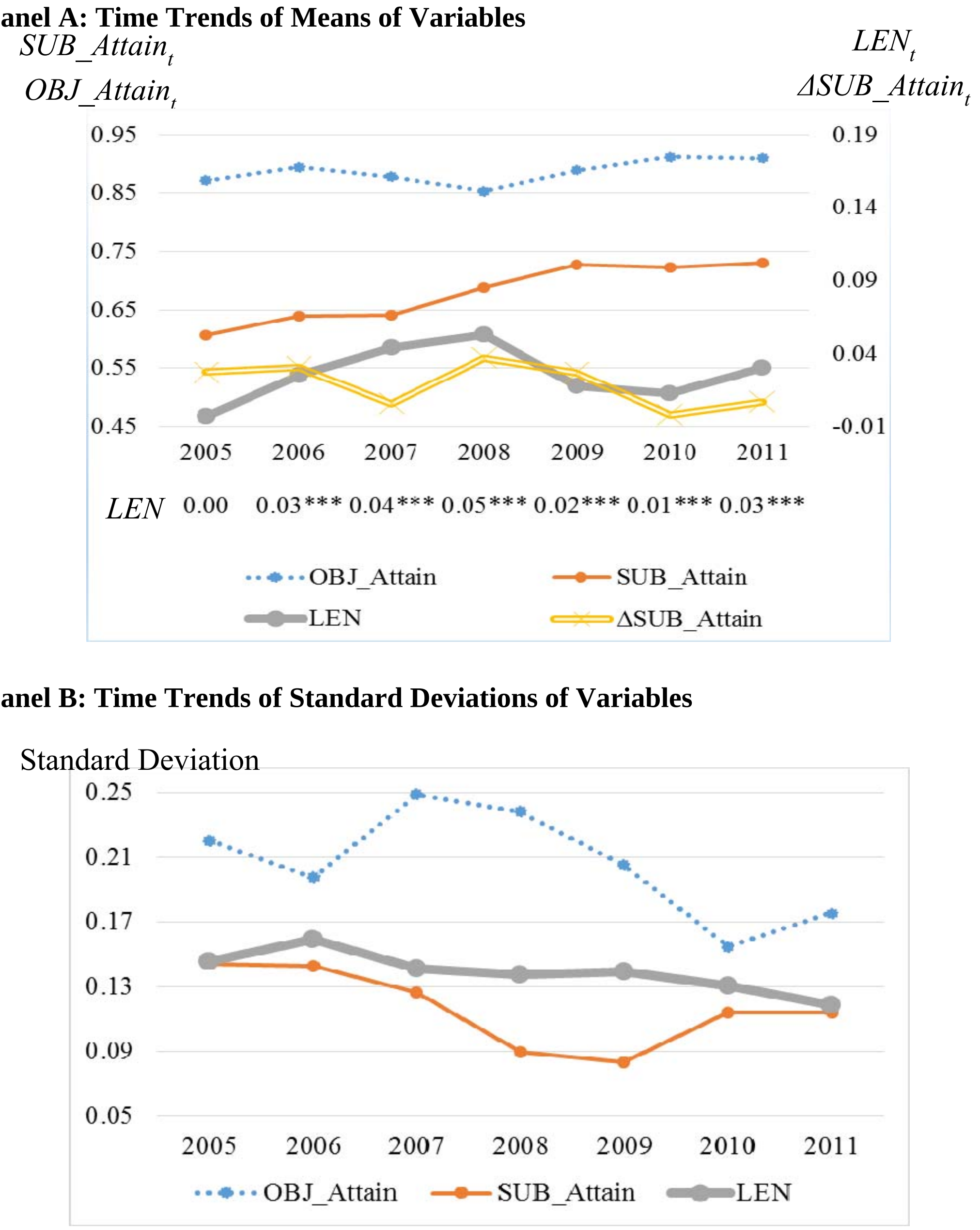


Panel B: Distribution of *OBJ_Attain*



Panel A shows the distribution of the score rate of subjective performance measure (*SUB_Attain*). The mean (median) value of subjective measure is 0.664 (0.650) and the standard deviation is 0.131. Panel B shows the distribution of score rate of objective performance measure (*OBJ_Attain*). The mean (median) value of objective measure is 0.880 (1.000) and the standard deviation is 0.205. Both panel have eight bins which have the interval of 0.125. The sample includes 7,470 subjective performance measures and 6,617 objective performance measures from 2005 to 2011.

FIGURE 2
Time Trends of Performance Score Rate and Leniency



Panel A: Time Trends of Means of Variables

Panel B: Time Trends of Standard Deviations of Variables

The sample includes 7,470 subjective performance measures (SUB_Attain_t), changes in subjective performance measures (ΔSUB_Attain_t), leniency measures (LEN_t) and 6,617 objective performance measures (OBJ_Attain_t) from 2005 to 2011 in both Panel A and Panel B.

TABLE 1
Examples of Performance Rating and Composition of Raters

Panel A: Example of Performance Ratings for an SOE in 2005[a]

Performance Category	Individual Measure	Measurement [b]	Commonality	Weight	Rating	Score	Attain (=Score rate)
1. Overall management	Capital productivity	OBJ	Common	5	-	3.998	0.800
	Customer satisfaction	OBJ	Common	2	-	1.512	0.756
	Restructuring or management innovation	SUB	Common	4	C	2.000	0.500
	Managing board of directors	SUB	Common	6.25	C	3.125	0.500
	…						
2. Main business	Maintaining high load factor	OBJ	Unique	3	-	3.000	1.000
	Effort for demand management	SUB	Unique	4	A0	3.500	0.875
	Efficiency of overseas business	SUB	Unique	2	B+	1.500	0.750
	…						
3. Business management	Financial structure	OBJ	Common	5	-	2.165	0.433
	Labor union management	SUB	Common	2	B+	1.500	0.750
	Budget management	SUB	Common	5	A0	4.375	0.875
	…						
SUB/C total				40		34.625	
SUB/U total				20		14.875	
OBJ/C total				23		16.869	
OBJ/U total				17		16.249	
Total				100		82.618	

Panel B: Example of Composition of Raters in 2008[c]

SOE Class	Subjective measure category			Objective measure	No. of raters
	Overall management	Main business	Business management		
Overall	3 professors				3
1. SOC (14 SOEs)	4 professors 2 industrial experts	5 professors 2 industrial experts	6 professors	2 professors 2 CPAs	23
2. Service (10 SOEs)	4 professors 1 industrial expert	4 professors 1 industrial expert	5 professors	1 professor 3 CPAs	19
3. Inspection management (6 SOEs)	3 professors	4 professors 1 industrial expert	3 professors	2 CPAs	13
4. Culture (9 SOEs)	3 professors 1 industrial expert	4 professors	3 professors 1 industrial expert	1 professor 2 CPAs	15
5. Industry promotion I (10 SOEs)	6 professors	4 professors	3 professors 1 CPA	3 CPAs	17

6. Industry promotion II (6 SOEs)	3 professors	3 professors	2 professors 1 industrial expert	1 professor 1 CPA	11
7. Education (6 SOEs)	2 professors 1 industrial expert	3 professors	2 professors 1 industrial expert	1 professor 2 CPAs	12
8. Pension (14 SOEs)	5 professors 1 industrial expert	5 professors 1 industrial expert	7 professors	4 CPAs	23
Total (75 SOEs)	33 professors 6 industrial experts	32 professors 5 industrial experts	31 professors 3 industrial experts 1 CPA	6 professors 19 CPAs	136

Panel A shows the example of performance ratings for KEPCO (Korea Electronic Power Corporation) in 2005. OBJ and SUB in the measurement colum in Panel A are objective and subjective performance measures, respectively. Panel B presents the example of raters composition in 2008.

TABLE 2
Descriptive Statistics of the Sample

Variables [a]	N	Mean	Standard Deviation	Min	Q1	Median	Q3	Max
LEN	7,470	0.025	0.144	-0.608	-0.067	0.024	0.117	0.644
SUB_Attain	7,470	0.664	0.131	0.125	0.600	0.650	0.750	1.000
Δ*SUB_Attain*	7,470	0.021	0.124	-0.550	-0.050	0.000	0.125	0.625
UNIQUE	7,470	0.296	0.456	0.000	0.000	0.000	1.000	1.000
WEIGHT	7,470	3.706	1.830	0.300	3.000	3.000	4.100	12.000
EXPERIENCE	7,470	2.107	0.376	1.364	1.829	2.125	2.250	3.222
WOMAN	7,470	0.075	0.046	0.000	0.045	0.073	0.104	0.167
Δ*OBJ_Attain*	7,470	0.003	0.100	-0.441	-0.051	0.004	0.051	0.388
SIZE	7,470	6.391	1.430	2.303	5.176	6.430	7.244	10.363
ROOM	7,470	2.350	0.662	0.000	1.951	2.402	2.808	3.920

The sample includes 7,470 unique measure-firm-years and 109 unique SOEs from 2005 to 2011. Data for performance rating and rater characteristics are manually collected from the performance evaluation report of SOEs (www.alio.go.kr).

a Variable Definitions:

LEN =	defined in the "RESEARCH DESIGN" section;
SUB_Attain =	performance score rate of a subjective measure;
Δ*SUB_Attain* =	change in subjective performance score rate (=SUB_Attain_{ijt} – $SUB\_Attain_{ij,t-1}$);
UNIQUE =	1 if the measure is a unique one, 0 if it is a common one;
WEIGHT =	weight assigned to the measure;
EXPERIENCE =	average tenure of raters within rating committee;
WOMAN =	fraction of female raters, which is the number of female raters within an SOE class divided by the number of total raters within an SOE class;
Δ*OBJ_Attain* =	change in sum of objective performance score rate (=OBJ_Attain_{jt} – $OBJ\_Attain_{j,t-1}$);
SIZE =	natural logarithm of the number of employees; and
ROOM =	room for improvement (= $\log_e$ (101 – $OBJ\_Attain_{j,t-1}$*100)).

where *i*, *j*, and *t* indicate performance measure *i*, firm *j*, and year *t*, respectively.

TABLE 3
Pearson Correlation Matrix

	LEN	*SUB_Attain*	*UNIQUE*	*WEIGHT*	*EXPERIENCE*	*WOMAN*	*SIZE*	*ROOM*
LEN	1.000							
SUB_Attain	0.592***	1.000						
	(0.00)							
UNIQUE	0.030**	0.060***	1.000					
	(0.01)	(0.00)						
WEIGHT	0.063***	0.081***	-0.185***	1.000				
	(0.00)	(0.00)	(0.00)					
EXPERIENCE	0.030**	0.138***	0.002	0.086***	1.000			
	(0.01)	(0.00)	(0.85)	(0.00)				
WOMAN	-0.026**	-0.015	-0.005	0.026**	-0.059***	1.000		
	(0.03)	(0.20)	(0.69)	(0.03)	(0.00)			
SIZE	0.068***	0.290***	0.005	-0.025**	0.133***	0.081***	1.000	
	(0.00)	(0.00)	(0.68)	(0.03)	(0.00)	(0.00)		
ROOM	0.346***	-0.087***	0.009	-0.021*	-0.037***	-0.139***	-0.019	1.000
	(0.00)	(0.00)	(0.45)	(0.07)	(0.00)	(0.00)	(0.10)	

P-values are reported in parentheses under each estimated correlation value. The symbols *, **, and *** correspond to 10 percent, 5 percent, and 1 percent significance levels, respectively, for two-tailed t-tests.

TABLE 4
Regressions of Leniency on Prior Performance Score Rate [a]

		(1)	(2)	(3)	(4)	(5)	(6)
Independent Variables [b]	Pred. Sign	Coeff. (t-value)	Coeff. (t-value)	Coeff. (t-value)	Coeff. (t-value)	Coeff. (t-value)	Coeff. (t-value)
SUB_Attain_{t-1}	–	-0.147***					
		(-6.79)					
$High\_SUB_{t-1}$	–		-0.035***	-0.050***			
			(-7.32)	(-10.44)			
Low_SUB_{t-1}	+		0.012**	0.023***			
			(2.48)	(4.83)			
LEN_{t-1}	–				-0.337***		
					(-24.55)		
$High\_LEN_{t-1}$	–					-0.070***	-0.078***
						(-14.49)	(-14.68)
Low_LEN_{t-1}	+					0.046***	0.052***
						(12.70)	(10.57)
$UNIQUE_t$	+/–	0.012***	0.011***	-0.120	0.016***	0.014***	0.075**
		(3.10)	(2.92)	(-1.54)	(3.29)	(3.09)	(2.35)
$WEIGHT_t$	+	0.007***	0.006***	0.007***	0.010***	0.009***	0.009***
		(3.07)	(3.11)	(5.40)	(4.21)	(4.52)	(5.58)
$EXPERIENCE_t$	–	-0.019***	-0.020***	-0.020***	-0.006	-0.009	-0.010
		(-2.61)	(-2.76)	(-3.18)	(-0.56)	(-0.90)	(-1.35)
$WOMAN_t$	+/–	-0.027	-0.030	-0.033	0.036	0.017	0.023
		(-0.65)	(-0.72)	(-0.60)	(0.84)	(0.36)	(0.38)
$SIZE_t$	+/–	0.003	0.004	0.001	-0.033***	-0.035***	-0.035*
		(0.26)	(0.32)	(0.04)	(-2.87)	(-2.89)	(-1.83)
Intercept		0.136	0.031	0.139	0.355***	0.385***	0.253
		(1.09)	(0.23)	(0.86)	(2.81)	(2.92)	(1.26)
Year Fixed Effects		Yes	Yes	Yes	Yes	Yes	Yes
SOE Fixed Effects		Yes	Yes	Yes	Yes	Yes	Yes
Measure Fixed Effects				Yes			Yes
Number of observations		7,470	7,470	7,470	4,936	4,936	4,936
Adjusted R^2		0.124	0.121	0.098	0.197	0.170	0.145

T-statistics are reported in parentheses under each estimated coefficient. Standard errors are corrected for heteroskedasticity using the Huber-White robust standard errors clustered by performance measure. The symbols *, **, and *** correspond to 10 percent, 5 percent, and 1 percent significance levels, respectively, for two-tailed t-tests. Please refer to the paper for a detailed explanation of these tests.

a This table shows the coefficient estimates of the leniency determinants by using the following equations:

$$LEN_{ijt} = \alpha_0 + \alpha_1\, SUB\_Attain_{ij,t-1}\ (or\ LEN_{ij,t-1}) + \alpha_2 UNIQUE_{ijt} + \alpha_3\, WEIGHT_{ijt}$$

$$+\alpha_4\, EXPERIENCE_{jt} + \alpha_5\, WOMAN_{jt} + \alpha_6\, SIZE_{jt} + Fixed\ Effects + \varepsilon_t. \qquad (1)$$

$$LEN_{ijt} = \alpha_0 + \alpha_1\, High\_SUB_{ij,t-1} (or\ High\_LEN_{ij,t-1}) + \alpha_2\, Low\_SUB_{ij,t-1} (or\ Low\_LEN_{ij,t-1})$$
$$+ \alpha_3\, UNIQUE_{ijt} + \alpha_4\, WEIGHT_{ijt} + \alpha_5\, EXPERIENCE_{jt} + \alpha_6\, WOMAN_{jt} + \alpha_7\, SIZE_{jt}$$
$$+ Fixed\ Effects + \varepsilon_t. \qquad (2)$$

where *i*, *j*, and *t* indicate performance measure *i*, firm *j*, and year *t*, respectively.

b The independent variable, Low_SUB_{t-1} ($High\_SUB_{t-1}$) is an indicator variable which refers to 1 if the score rate belongs to the lowest (highest) quartile of *SUB_Attain* in the previous year, 0 otherwise. Low_LEN_{t-1} ($High\_LEN_{t-1}$) is an indicator variable which refers to 1 if the leniency tendency belongs to the lowest (highest) quartile of *LEN* in the previous year, 0 otherwise. See Table 2 for the definitions of other variables.

TABLE 5
Regressions of Alternative Measures of Leniency on Prior Performance Score Rate[a]

		Dependent Variable [b]			
Independent Variables	Predicted Sign	(1) $LEN2_t$	(2) $LEN3_t$	(3) ΔSUB_Attain_t	(4) $\Delta SUB\_Attain2_t$
$High\_SUB_{t-1}$	–	-0.035***	-0.030***	-0.114***	-0.038***
		(-7.34)	(-6.21)	(-18.36)	(-6.70)
Low_SUB_{t-1}	+	0.012**	0.023***	0.100***	0.025***
		(2.55)	(6.72)	(28.23)	(5.31)
$UNIQUE_t$	+/–	0.011***	0.010***	0.010***	0.015***
		(2.96)	(2.78)	(2.81)	(2.69)
$WEIGHT_t$	+	0.006***	0.007***	0.006***	0.004
		(3.11)	(3.15)	(3.46)	(1.57)
$EXPERIENCE_t$	–	-0.020***	0.009	0.006	0.010
		(-2.78)	(1.57)	(1.13)	(1.32)
$WOMAN_t$	+/–	-0.029	-0.087***	-0.084**	-0.059
		(-0.69)	(-2.67)	(-2.43)	(-1.45)
$SIZE_t$	+/–	0.004	0.011	0.012	-0.042
		(0.29)	(0.88)	(0.88)	(-1.38)
Intercept		0.001	-0.080	-0.088	0.467
		(0.01)	(-0.63)	(-0.64)	(1.50)
Year Fixed Effects		Yes	Yes	Yes	Yes
SOE Fixed Effects		Yes	Yes	Yes	Yes
Number of observations		7,470	7,470	7,470	3,310
Adjusted R^2		0.130	0.118	0.299	0.129

T-statistics are reported in parentheses under each estimated coefficient. Standard errors are corrected for heteroskedasticity using the Huber-White robust standard errors clustered by performance measure. The symbols *, **, and *** correspond to 10 percent, 5 percent, and 1 percent significance levels, respectively, for two-tailed t-tests. Please refer to the paper for a detailed explanation of these tests.

a This table estimates the impact of prior performance score rate on leniency by using the following equation:

$$LEN_{ijt} = \alpha_0 + \alpha_1 High\_SUB_{ij,t-1} + \alpha_2 Low\_SUB_{ij,t-1} + \alpha_3 UNIQUE_{ijt} + \alpha_4 WEIGHT_{ijt} + \alpha_5 EXPERIENCE_{jt} + \alpha_6 WOMAN_{jt} + \alpha_7 SIZE_{jt} + Fixed\ Effects + \varepsilon_t. \quad (2)$$

where *i*, *j*, and *t* indicate performance measure *i*, firm *j*, and year *t*, respectively.

b The dependent variable, $LEN2_t$ is calculated as subjective performance score rate (SUB_Attain_t) minus modified benchmark ($BENCHMARK2_t$) minus change in objective performance score (ΔOBJ_Attain_t). The modified benchmark ($BENCHMARK2_t$) is driven by the distribution of new measures which is a substitute of forced distribution assumption. $LEN3_t$ is LEN_t plus change in objective performance score

(ΔOBJ_Attain_t). Change in subjective performance score rate (ΔSUB_Attain_t) is the difference between SUB_Attain_{ijt} and $SUB\_Attain_{ij,t-1}$. The dependent variable in column (4), $\Delta SUB\_Attain2_t$ is the gap between SUB_Attain_{ijt} and average subjective performance score rate of prior three years. See Table 2 for the definitions of other variables.

Essay2. Determinants of Narrative Feedback in Subjective Performance Evaluation

I. INTRODUCTION

Subjective performance evaluation complements and corrects distorted objective measure (Banker and Datar 1989; Baker et al. 1994; Bol 2008; Bushman et al. 1996; Gibbs et al. 2004; Holmström and Milgrom 1991; Murphy 1999) and provides feedback to the subordinate about his performance, which helps him to better allocate his efforts (Brutus 2010; Murphy and Cleveland 1995). Performance evaluation usually combines a rating scale and written comments about the results of and suggestions to evaluatees' performance. In other words, performance evaluation systems include a qualitative component in the form of narrative feedback which is not limited to rating scales (Smither and Walker 2004; Spence and Keeping 2011). However, most prior studies in performance evaluation have focused on the use of numerical ratings and rarely investigated how evaluators give narrative feedbacks to subordinates and how the subordinates react to it.

Specifically, accounting research has found that corporate narrative reporting constitutes an important tools of communication with firm's stakeholders (Li 2010). Narrative disclosures contain information about the data generating function of numeric financial data and reveal the managers' incentive of firm's decision making (Li 2011; Simon 1997). Similarly, in performance evaluation area, anecdotal and empirical evidence indicates that feedback recipients pay a great deal of attention to evaluators' narrative comments (Smither and Walker 2004; Stetz and Ford 2010). It is clear that numerical ratings do not well capture the nuances and subtleties that written narratives can (Toegel and Conger 2003). Subordinates have a strong need for narrative feedback because it helps them interpret the numerical ratings (Antonioni 1996), and provide specific appreciations and suggestions for future development (Brutus 2010). However, only recently has some research started to examine the role of narrative formats in performance evaluation as technological advances make it possible to process the narrative information (Atwater and Brett 2006; Brutus et al. 2013; Smither and Walker 2004).

Understanding the causes and performance consequences of narrative feedback is important for both evaluators and evaluatees, as it can help them determine the incentives surrounding the performance evaluation system and can lead to improved incentive contracting. Prior empirical studies mainly focus on the consequence part of narrative feedback in performance evaluation. Atwater and Brett (2006) compare feedback formats (text versus number), and find that numerical feedback has more positive impact on subordinates reaction because rating number is more specific and comparable information than narrative feedback. These findings are opposite to their prediction which suggests that text format increases acceptance of feedback (Ilgen and Davis 2000). Smither and Walker (2004) empirically find that amount, nuance, and specificity of narrative feedback affect the improvement of managers. In short, empirical evidence on the determinants of narrative feedback is scarce (Harackiewicz and Larson 1986; Larson 1984).

This study extends the current literature by focusing on what causes evaluators to present as many or as few comments or suggestions. To address the questions, I examine the narrative feedback of Korean State-Owned Enterprises (hereafter, SOEs). In particular, I manually collect the narrative feedback data of 17 SOEs for the period from 2004 to 2011, and investigate the determinants of narrative feedback amount. More specifically, I first divide total words amount of narrative feedback to those of including suggestions and others. Suggestion feedback might be more useful because it provides direction for improvement. This suggestion conveys to evaluatees what the evaluator would like to see and why taking this action is important (Cannon and Whitherspoon 2005). Then, I examine whether extrinsic motivation of evaluators – information acquisition cost and confrontation cost – and intrinsic one increase the amount of total comment or suggestion, by utilizing my rich data.

The evaluation format will have a significant influence on the cognitive processes underlying the evaluation process. The use of numeric ratings is analogous to a matching task, one for which the evaluators are asked to find an appropriate equivalent for their evaluations of the target's performance (Borman 1991). Narrative comment, however, provides little guidance to the evaluators and hinges on cognitive activities which need to communicate their evaluation in writing (e.g., Fitzer 2003; Flower and Hayes 1981; Kellog 1994). Putting one's thought into words is a difficult task, one that requires precise and specific information. Moreover, creating suggestions for improvement also increases the cognitive demands of evaluators. The element of suggestion for improvement is not evaluative in nature, rather it requires a solution to the performance weaknesses diagnosed (Brutus 2010). So, I argue that evaluators tend to conduct incomplete evaluation including giving fewer narrative feedbacks when they face high information acquisition costs in order to minimize the time and effort invested in the performance evaluation process.

Evaluators also have their own incentives to give narrative feedback. Evaluators are likely to avoid confrontations to subordinates and limit criticism from other related parties (Harris 1994). Regarding rating score, when the performance evaluation systems are based on a combination of numerical ratings and narrative comments, evaluators might produce narrative comments that have a certain level of consistency with ratings (Smither and Walker 2004). Also, I hypothesize that evaluators are likely to give more narratives about the performance measure for which subordinates care much. These narrative comments can be seen as attempts to help the evaluators avoid and mitigate confrontation costs by explaining and justifying the provided performance rating, especially highly concerned measures by subordinates. Also, there exists accountability issue to one's supervisors. Evaluators' supervisors usually authorize the final ratings and feedback contents (Bernardin and Villanova 1986). A form requiring extensive documentation about the rating score increases likelihood of supervisor criticism if the evaluator does not engage in the necessary information processing activities (Harris 1994). This will increase evaluator motivation of providing more narratives. Specifically, feedback enhancing behavior evinced by supervisors may be more influenced by how their behavior will look to their supervisors than by the reaction from the subordinates (Ferris and Judge 1991). In sum, I predict that confrontation cost from both subordinates and supervisors will increase the amount of total comment and suggestion.

Motivated evaluators tend to help subordinates improve their performance. Evaluators tend to value attainment of intrinsic rewards from engaging in performance evaluation activities. Proving helpful feedback to subordinates results in increased esteem and recognition from subordinates and supervisors (Harris 1994). Therefore, evaluators' motivation is likely to affect the thoroughness of the integration process of information and giving narrative feedback. Accordingly, I hypothesize that a motivated evaluators will be more likely to provide more narrative feedback including suggestion for improvement.

The results show that evaluators respond to their own interest and preferences when they provide the narrative feedback. My findings indicate that (1) information acquisition costs negatively affect the amount of comments and suggestions, and (2) confrontation cost and intrinsic motivation of evaluators are positively associated with the amount of comments and suggestions. More specifically, I find that evaluators present large amount of narrative feedbacks about the highly weighted, long–tenured, and both high scored and low scored performance measures, and SOEs with large number of evaluators and large number of employees. More importantly, the determinants about the ratio of narrative suggestions to total narrative comments are negatively associated with performance score, measurement uniqueness and evaluator group size, and positively associated with evaluators' experience, age of performance measure and woman fraction within evaluator group.

My study contributes to the literature in the following way. This study answers several calls for empirical evidence on the determinants of narrative comments amount and suggestion amount (Brutus 2010; Spence and Keeping 2011). Factors affecting motivation of evaluators can be classified to extrinsic and intrinsic ones (Harris 1994; Murphy and Cleveland 1995). An important extrinsic determinants of evaluator behavior is avoidance of negative consequences (Longenecker et al. 1987), such as information acquisition cost and confrontation cost. I present that evaluators' extrinsic and intrinsic motivations are the crucial determinants of narrative feedback. My analysis shows that the narrative comments are affected and bounded by the determinants which are related to information acquisition costs, confrontation costs and intrinsic willingness to influence, even though narrative comments are theoretically unbound. This finding indicates that subjective performance evaluation is explained by managers' incentives and preferences (Bol 2011; Harris 1994; Kane 1994; Woods 2012).

The remainder of the paper is organized as follows. I review the theoretical background and develop my hypotheses in Section II. In Section III, I provide institutional background on performance evaluation systems and performance feedback of SOEs in Korea. In Section IV, I describe research design including sample selection and empirical measures. My primary empirical results are provided in Section V. Section VI concludes.

II. THEORY AND HYPOTHESIS DEVELOPMENT

Giving subordinates feedback on how well they are doing in their jobs is held to meet a variety of needs. From the evaluator's point of view, it assists effective learning so that tasks are completed correctly and helps maintain and stimulate effort towards desired goals. Also, from the subordinate's viewpoint, feedback can satisfy any personal need for information on progress and facilitate social comparison with others (Fletcher 1986). In organizational practice, performance ratings are often supplemented with qualitative parts for providing comments or with comment-intensive developmental feedback processes. Subordinates may receive comments from their evaluator as part of their performance evaluation (Brutus 2010; Dalessio 1998; Harrington 2012). However, performance evaluation literature has focused mainly on the characteristics of numerical ratings, not on the use or quality of narrative comments (Brutus 2010). For example, researchers have examined properties of rating scales (Landy and Farr 1980), and biases in ratings (Ahn and Hyun 2014; Bol 2011; Bretz et al. 1992; Merchant et al. 2010; Moers 2005). However, given the extensive use of comments in performance evaluation practice (Atwater and Brett 2006; Brutus 2010), relatively few have investigated how qualitative comments are provided and how the subordinates react to them.

Cognitive psychologists, educational psychologists and linguists have long established the function of narratives in communication (e.g., Butler 1987; Chomsky 1986). Some researchers claim that narratives are more comprehensible and thus, more useful (Moxey and Sanford 2000; Sanford et al. 2002). It is clear that subordinates pay more attention to narrative comments than to numerical ratings they receive (Ferstl and Bruskiewicz 2000) because numerical ratings cannot capture the nuances and subtleties that written narratives can (Toegel and Conger 2003). Thus, subordinates have a strong need for narrative feedback because it helps them interpret the numerical ratings (Antonioni 1996), and provide specific guidance for improving their performance

(Brutus 2010; Cannon and Witherspoon 2005). In the past few years, researchers began investigating this underrepresented area in the performance evaluation literature. For example, Atwater and Brett (2006) compare feedback formats (text versus number), and argue that employees react to text feedback more favorably. However, they find that numerical feedback has more positive impact on subordinates' reaction which is contrary to their prediction. While they interpret the result in a way that rating number is more specific and comparable information than narrative feedback and hence, subordinates react to rating number more, it is unclear how the lack of rating number can create ambiguity in performance reactions of subordinates. Also, Smither and Walker (2004) empirically examine that performance improvement would be related to the amount of narrative comments each subordinate received, whether those comments are favorable or unfavorable, and whether the comments are behavior focused or trait focused. They show that 4 percent of the performance improvement can be explained by the favorability of comments and the overall number of comments. They also find that subordinates who received a large number of unfavorable, behavior focused comments underperform other subordinates.

Although these prior studies in performance evaluation document the consequence of narrative feedbacks, empirical evidence on the determinants of narrative feedback is scarce (Harackiewicz and Larson 1986; Larson 1984).

2.1. Determinants of Narrative Feedback Amount

2.1.1. Information Acquisition Cost

Narrative comments represent a much more expansive format of performance feedback than performance ratings. They provide an almost infinite range of options to evaluators in formulating their message. However, the structure of narrative comments provides little guidance to the evaluators and hinge on cognitive activities which need to communicate their evaluation in writing (e.g., Fitzer 2003; Flower and Hayes 1981; Kellog 1994). Also, writing narratives is a dynamic and complex process that draws from deep-seated knowledge (Flower and Hayes 1981). Therefore, evaluators have incentives to minimize the time and effort invested in the performance evaluation process, and these preferences will influence the evaluators' feedback behavior (Harris 1994). To present narrative feedback, especially if the evaluators suggest the narratives about subordinates' improvement (e.g., suggestion for development), managers need to invest time and effort in processing information on subordinate performance because writing comments requires more thought than the matching of ratings to behaviors (Harrington 2012). The demands related to the lack of directive guideline, the need to convey the evaluation in writing, and the possibility of providing suggestions for development will enhance the role played by memory-based processing in the production of narrative comments. Since evaluators have a preference to limit their time and effort spent on performance evaluation, narrative feedback will likely be insufficient when information acquisition costs are high. I then set up the hypothesis as follows.

H1a: Information acquisition cost is negatively associated with the amount of narrative feedback.

H1b: Information acquisition cost is negatively associated with the amount of suggestion for development.

2.1.2. Confrontation Costs

The incentive factors of evaluators underlying the use of comments are important in performance feedback. The motivation for evaluators has a significant influence on their production of comments (Brutus 2010). Evaluators are likely to avoid confrontations with subordinates and limit criticism from other interested parties (Harris 1994). Regarding rating score, when subordinates are unsatisfied with their rating, they will likely ask their evaluator for a justification of the performance rating they have received. Thus if the evaluator intends to deflate the ratings, or rate lower than the evaluatees would expect, the evaluators are motivated to obtain as much information as possible, and are prepared to produce that evidence as needed (Schmitt et al. 1991). Likewise, I predict that evaluators are more likely to give large amount of narratives about the performance measure in which subordinates care much. Acceptability of subordinates, rather than accuracy, is one of the most important criteria in a developmental context.[24] In doing so, evaluators are likely to feel that they are being kind to their subordinates and as a result expect to get exemption from criticism (Barrett 1966). In addition, narrative comments normally include the words of suggestion for development. Gillipsie et al. (2006) reported that nearly a third of narrative comments contain specific recommendations for the subordinates. Detailed suggestions for improvement might positively affect the subordinates' perception of fairness and make evaluators better prepared and able to engage in following discussions about that feedback. Thus, by providing suggestions for improvement, evaluators reduce confrontation burden and fell

[24] Most of the literature on how subordinates respond to performance information rests on the perception of performance evaluation systems and, more specifically, on perceptions of their accuracy, justice, and fairness (Gilliland and Langdon 1998; Keeping and Levy 2000; Smither 1998).

comfortable. These suggestions can be seen as attempts to have the effect of counterbalancing the negative evaluations.

Also, there exists accountability issue to one's supervisors. Evaluators' supervisors usually authorize the final ratings and feedback contents (Bernardin and Villanova 1986). After evaluators have subjectively rated subordinates and given feedback narratives, supervisors convene to review the subjective ratings and narratives before the performance reports are disseminated back to the subordinates (Sedatole and Woods 2013). A form requiring extensive documentation about the rating score increases likelihood of supervisor criticism if the evaluator does not engage in the necessary information processing activities (Harris 1994). This will increase evaluator motivation of providing more narratives. Specifically, feedback enhancing behavior evinced by supervisors may be more influenced by how their behavior will look to their supervisors than by the reaction from the subordinates (Ferris and Judge 1991).

In sum, these narrative comments can be seen as attempts to help the evaluators avoid and mitigate confrontation costs by explaining and justifying the provided performance rating, especially highly concerned measures by subordinates and supervisors.

H2a: Confrontation cost is positively associated with the amount of narrative feedback.

H2b: Confrontation cost is positively associated with the amount of suggestion for development.

2.1.3. Intrinsic Motivation of Evaluators

Performance feedbacks aim to develop self-awareness and persuade subordinates to increase their effort, change the direction of their effort, and persist in trying to reach their goal (Kanfer 1990). Motivated evaluators likely write more in order to attain the above targets because narrative comment is better equipped than standardized rating to capture individual performance. Specifically, suggestion for development is a direct vehicle to help inform development plans and understand the core message within the feedback. In other words, this knowledge effect of narrative comments might mainly come from the suggestions for development. In the presence of prescriptive information, recipients are more likely to adjust their strategies accordingly and work smarter (Audia and Locke 2003). Also, evaluators have incentives to help subordinates improve their performance because evaluators might expect to receive benefits back from them (Emerson 1976). This social exchange theory posits that this reciprocation effect could be at play with the use of narrative comments. Subordinates that receive lengthy narrative evaluations may recognize the efforts deployed by the evaluators and may be more likely to respond favorably, by instigating behavior changes.

In sum, I predict that when evaluators face high confrontation cost and are highly motivated, evaluators give large amount of narrative comments.

H3a: Motivation of evaluators is positively associated with the amount of narrative feedback.

H3b: Motivation of evaluators is positively associated with the amount of suggestion for development.

III. INSTITUTIONAL BACKGROUND

3.1. Overview of Performance Evaluation Systems of SOEs in Korea

The Korean government enacted the Law for Management of SOE in 2003 that requires SOE be evaluated annually by a group of auditors assigned by Ministry of Strategy and Finance and disclose the result of performance evaluation by June of the subsequent year. The government develops performance measures with which the evaluator and the evaluatee mutually agree, and distribute the guideline by the end of the year for evaluation. Then, at the end of the March, SOEs submit reports for the previous year's performance. From April to June, evaluators evaluate the performance of the SOE.

The performance evaluation system employs an incentive bonus plan in which bonus size is based on the results of performance evaluation. The bonus size is limited and determined by ranking-based peer comparison via distribution analysis.[25] Thus, it is important to compare relative scores among peers. To provide evaluatees (SOEs) with appropriate incentives and sound evaluation systems, the Korean government uses various measures under three categories: overall management, main business, and business management. Panel A in Table 1 presents an example of a performance rating of an SOE in 2005.

The third column of Panel A in Table 1 shows that each evaluation criterion classifies performance measures into objective and subjective measures. Objective measures are assessed using four quantitative, formula-based methods: actual-to-target analysis, target-range assignment analysis, trend analysis, and beta analysis.[26] In contrast

[25] After all the evaluations, the government transforms total performance score of SOEs into standardized Z-score. If Z-score of an SOE is above 2σ, then the SOE is graded 'S' and paid the maximum 500% bonus of monthly salary.

[26] (1) Actual-to-target analysis represents actual performance divided by target performance. Sales volume, labor expenses, and plant construction progress are examples of measures that use the actual-to-target method. (2) Target-range assignment analysis uses the ratio of actual performance minus minimum target performance to maximum target, minus minimum target performance. Capital productivity, capacity utilization, and customer satisfaction index use this evaluation method. (3) Trend analysis is a regression

to objective measures, subjective measures have a single evaluation method, that is, grading.[27] While objective measures are calculated by comparing actual performance based on predetermined formula for benchmark performance, assigning subjective measures to the various grade levels depends entirely on the evaluator's subjective evaluation.

For only subjective performance measures, after rating decision, it is necessary to write down the narrative comments which include the reason for the ratings and the suggestion for development.[28] There is no guideline about the amount or characteristics about narrative comments. After initial performance evaluation, evaluatees have opportunities to express their opinions about the processes and the results of evaluation which include ratings and narrative comments. Also, the supervisor (Korean government) and evaluators (evaluating committee) discuss and investigate the soundness of evaluation results together. Finally, inspection institute of Korean

analysis that computes standard performance coefficients using past actual performance (e.g., the prior 15 years' figures). Actual performance is evaluated against expected performance via the standard performance coefficient. This method is, in general, used to assess how effectively SOEs manage their inventory, cost of capital, cost of goods sold, and administration expenses. (4) Beta analysis is similar to trend analysis in using past data to obtain a benchmark to assess current performance; however, it uses the beta distribution instead of the regression. Typical measures that use this method are labor productivity, economic value added (EVA), and plant power management. After conducting the quantitative assessment, each SOE is assigned into five or nine grades, in accordance with the predetermined score ranges.

[27] Subjective measures complement objective measures for activities that are difficult to quantify, but that are important to achieving firm strategic goals—for example, efficient management in strategic plans, improvement in control systems, appropriateness in assessment and implementation of investments, development of organizational culture, appropriateness of budgeting, cooperation with community, and employee education.

[28] Regarding the hypothesis, information acquisition cost, confrontation cost and intrinsic motivation can be applied to rating behavior (Bol 2011; Murphy and Cleveland 1995) as well as providing narratives. However, motivations of rating are different from those of writing narratives because of their characteristics. As stated above, writing narratives is a complex process that draws from deep-seated knowledge (Flower and Hayes 1981), and feedback recipients pay a great deal of attention to evaluators' narrative comments (Smither and Walker 2004; Stetz and Ford 2010). Thus, both information acquisition cost and confrontation cost are greater when evaluator provide narratives.

government carries out its duties as the watchdog of the appropriateness of evaluation processes and results.

3.2. Formation of Evaluators

All SOEs are categorized into 8 classes based on their business characteristics – SOC, Service, Inspection management, Culture, Industry promotion I, Industry promotion II, Education, and Pension, as shown in Panel B in Table 1.[29] Each class is audited by four groups of evaluating committee – one objective measure team and three subjective measure teams that grade overall management, main business, and business management. For performance evaluation of SOEs, evaluating committees consist mainly of academic professors, certified public accountants, and industrial experts. They should have no private connection to the matched SOEs, and hence this strict requirement of independence from the evaluatees might ensure fair evaluation. Evaluators are allocated with matrix structure in order to harmonize between class characteristics and measurement characteristics. For a measurement group of common measures (overall management and business management), evaluators consist of experts in organization management such as accounting, public administration, business, and economics. On the other hand, for a measurement group of unique measures (main business), evaluators are mostly professionals with experiences in the evaluatees' industries.

[INSERT TABLE 1 ABOUT HERE]

The government also provides various safeguards against evaluator bias. First, if a

[29] Panel B in Table 1 provides an example of composition of evaluators in 2008. The number of classification category varies six to nine as the rule changes.

evaluator is assigned to a certain evaluating committee in the current year, the evaluator would usually move to a different committee every year to promote independence (i.e. rotation system). Second, after three years of rating service, the evaluator should take one year's leave of absence. Moreover, evaluators should sign a code of ethics in evaluation and receive a training about how to deal with interest conflicts. All these rules guarantee independence and fairness of evaluation process and reduce familiarity bias.

IV. RESEARCH DESIGN

4.1. Sample Selection

My sample is manually collected from the Korean Government's annual reports on its performance evaluations for SOEs from 2004 to 2011.[30] I obtain a final sample of 2,120 measure-SOE-year observations, representing 17 Korean SOEs for sample period of eight years. To analyze the narrative comments in the performance report, two research assistants independently read the every performance evaluation report and classify the words into different categories based on whether the words are related with the suggestion for development.

4.2. Regression Model

[30] The Korean government enacted the Law for Management of SOEs in 1984. External evaluators have evaluated the performance of 13 SOEs since 1985. However, this data of performance results makes it difficult to compare inter-temporal rating tendency because (1) before and after year 2004, there is a big difference in rating systems that can mislead the effects of leniency over time, (2) in 1998, the minimum base score at 75% was eliminated, and hence there is a large difference in score rates between before-1998 and after-1998 (Ahn et al. 2010).

I develop a cross-sectional model for determinants of the amount of narrative feedback and estimate Equations (1A), (1B) and (2) using OLS to test if the information acquisition costs, the confrontation costs and motivation of suggestion for development affects the evaluators' likelihood of providing narrative feedback:

$$\begin{aligned}&ln(\#Total{+}1)_{ijt} = \alpha_0 + \alpha_1 Attain_{ijt} + \alpha_2 Unique_{ijt} + \alpha_3 Weight_{ijt} + \alpha_4 Measure\_Age_{ijt}\\&+\alpha_5 Size_{jt} + \alpha_6 Experience_{jt} + \alpha_7\, \#Evaluator_{jt} + \alpha_8\, Expert_{jt} + \alpha_9\, Woman_{jt}\\&+ Fixed\ Effects + \varepsilon_{ijt}. \qquad (1A)\end{aligned}$$

$$\begin{aligned}&ln(\#Suggest{+}1)_{ijt} = \alpha_0 + \alpha_1 Attain_{ijt} + \alpha_2 Unique_{ijt} + \alpha_3 Weight_{ijt}\\&+ \alpha_4 Measure\_Age_{ijt} + \alpha_5 Size_{jt} + \alpha_6 Experience_{jt} + \alpha_7\, \#Evaluator_{jt} + \alpha_8\, Expert_{jt}\\&+ \alpha_9\, Woman_{jt} + Fixed\ Effects + \varepsilon_{ijt}. \qquad (1B)\end{aligned}$$

$$\begin{aligned}&Ratio_{ijt} = \alpha_0 + \alpha_1 Attain_{ijt} + \alpha_2 Unique_{ijt} + \alpha_3 Weight_{ijt} + \alpha_4 Measure\_Age_{ijt}\\&+\alpha_5 Size_{jt} + \alpha_6 Experience_{jt} + \alpha_7\, \#Evaluator_{jt} + \alpha_8\, Expert_{jt} + \alpha_9\, Woman_{jt}\\&+ Fixed\ Effects + \varepsilon_{ijt}. \qquad (2)\end{aligned}$$

where *i*, *j*, and *t* indicate performance measure *i*, firm *j*, and year *t*, respectively.

All regressions in this paper are estimated with Huber-White robust standard errors clustered by year and performance measure level. The standard errors are robust to both serial correlation and cross-sectional correlation (Gow et al. 2010). I build upon related literature on both evaluators motivation for evaluation and feedback (Bol 2011; Brutus 2010), in order to construct the measures used in my empirical tests. Below I provide a definition of my proxy for the three hypothesis and a description of the factors that prior research suggests could affect the amount of narrative feedback at firm-level, evaluator-level, and individual measurement-level.

4.2.1. Dependent Variables

Performance feedback can be interpreted as hiving both backward-directed and forward-directed loops in terms of time (Otely 1999; Pitkänen and Lukka 2011). While backward-directed feedback provides comparative information between actual performance and predetermined performance target, feed-forward information forecasts the need for actions of planning and improvement. Nishmura (2003) argues that it could be useful to view feed-forward information with separate concept. Following this classification, I use two types of narrative comments as a dependent variable – the number of narrative comments and the number of suggestion words in this paper.

For the coding of narrative comments, it is important to specify a unit of analysis. A sentence or paragraph can include multiple themes, and multiple sentences or paragraphs can be articulated around a single performance-related theme (Brutus et al. 2013; Smither and Walker 2004). I alternatively regard number of words as a narrative unit because it provides objective and reliable information about coding process.[31] The coding process is exemplified in Appendix A. First, I divide total narrative comments of a performance measure by paragraph. This generates six paragraphs in the example. The sixth paragraph in the example indicates just the rating of the performance measure and hence, it has little meaning with respect to feedback effect. So, I exclude it when I count the number of words. Second, I count the number of words each paragraph has. The "#Comment" column presents the total number of words in the paragraph. Also, if the paragraph includes more than one suggestion, I count total number of words in the paragraph and regard it as the number of suggestion words. Finally, I can get both the

[31] To check the robustness of our coding process, I also analyze the determinants by using the alternative narrative unit – number of paragraph, and find similar results to the main findings. I present the detailed results on the additional analysis part in the Empirical Results section.

total number of comments and the number of suggestion for development at the performance measure level.

4.2.2. Determinants Influencing the Amount of Feedback

The model includes a comprehensive set of determinant variables which include characteristics of evaluators, SOEs (evaluatees), and performance measures. As stated in the Hypothesis Development section, I predict that the information acquisition costs, the confrontation costs, and intrinsic motivation can affect the amount of narrative comments.

Performance score rate ($Attain_t$) is an important factor in influencing the amount of narrative comments. Evaluators use comments to explain or elaborate on numerical ratings (Dalessio 1998). For example, if an evaluatee is rated low at a performance measure, the evaluatee can be longing for the reason of low rating. Evaluators can provide text to the evaluatee to reduce confrontation burden by commenting on specific instances in which the evaluatee performs poorly and on suggestions for development. Also, supervisors can have much attention to the reason for high ratings. In this case, evaluators feel pressure to justify their rating by providing sufficient comments. So, with regard to confrontation costs, I expect that performance score level has U-shaped association with the amount of narrative comments and negative association with the amount of suggestion for development. Furthermore, motivating low performers is considered an essential role of performance evaluation (Murphy and Cleveland 1995). Narrative comments may help feedback recipients learn about the way of improving performance. Specifically, the evaluatee who delivers a poorly structured behavior knows what areas of behavior she needs to improve and addresses the deficiency if the suggestion for development is given (Ghorpade 2000; Smither and Walker 2004). Therefore, to benefit back from the low evaluatees who have high probability of improvement, evaluators are more likely to provide narrative comments, especially for suggestion comments. Also, with respect to evaluators, it is relatively easier to provide suggestion comments for low performers because they have much rooms for improvement and hence, their weaknesses and things to consult and suggest for improvement can be easily found. Thus, both information acquisition costs and motivation of providing comments can negatively affect the total amount of narrative comments and the suggestion comments. The test variables of interest are the performance score level – $Attain_t$, Low_Attain_t and $High\_Attain_t$. The independent variable, Low_Attain_t ($High\_Attain_t$), is an indicator variable that equals to one if the score rate belongs to the lowest (highest) quartile of *Attain* in the current year, and zero otherwise.

Unique is an indicator variable that equals one if a performance metric is unique to a SOE, and zero if it is common to multiple SOEs. Unique meatrics are cognitively more difficult to measure than common meatrics (Lipe and Salterio 2000). In addition, some dimensions of performance in public organizations are often hard to measure (Burgress and Ratto 2003). Therefore, evaluators are less likely to write comment for unique measures that incur high information acquisition costs. On the other hand, because of high comparability of common measures among evaluatees, evaluators are more pressured to provide narrative comments and suggestions (Fried et al. 1999). Hence, I predict that the number of suggestion comments is negatively associated with uniqueness of the measure.

I use the weight assigned to a measure (*Weight*) to proxy for the relative importance among performance measures. Evaluators are more pressured to provide comments of highly weighted measures due to higher confrontation costs. I expect measurement weight to be positively associated with the amount of narrative comments and suggestion comments. The performance measure age (*Measure_Age*) can be defined as the number of years each performance measure has been with the SOE. As age of performance measure increases, the historical information is accumulated and evaluators can easily refer to the prior performance information. It is expected that the performance measure age positively affects the amount of narrative comments and suggestion comments due to the decreased information acquisition cost. *SIZE* is included because most previous studies support that evaluators may face larger confrontation costs from larger firms. Thus, large SOEs are likely to get larger amount of narrative comments than small SOEs do. Also, SOEs with more employees receive more public attention and media coverage (Du et al. 2012), and hence, evaluators and the government are more cautious when they evaluate and provide comments about the performance of large SOEs. These findings of prior studies let me predict the positive effect of firm size on the amount of narrative comments and suggestion comments.

For evaluator characteristics, I first proxy for evaluators' expertise and experience by the fraction of industrial experts within an evaluating committee, and the experience of evaluators which is measured by the average tenure of evaluators within an evaluating committee, respectively. Research in psychology also shows that greater degrees of expertise and experience evaluators make evaluation of evaluatees more reliable and more precise (Brown 1968; Schneier 1977). Therefore, as tenure of evaluators increases, and more industrial experts involves the evaluating committee, it is expected that the narrative comments contain lengthy and specific recommendation and suggestion for improvement. With respect to size of evaluating committee (*#Evaluator*), smaller teams tend to be more cohesive than larger teams because it is easier for a few people to agree on goals and coordinate work activities. Team members are reluctant to mention ideas because they believe that other team members silently evaluate them and teams take longer than individuals to make decisions. This is most common in meetings attended by people with different levels of status or expertise like my setting (McShane and Von Glinow 2008). Thus, an evaluator within a large team can be demotivated to provide sufficient narrative comments. This story leads me to predict the negative association between the number of evaluators in an evaluating committee and the amount of narrative comments and suggestion comments. *WOMAN* measures the gender effect of evaluators as the number of female evaluators divided by total number of evaluators within a class. A large body of field and laboratory work suggests that there may be gender-based differences in managerial behavioral tendencies. Specifically, women are likely to encourage subordinates by keeping open communication channels with their subordinates and sharing of task-relevant information (Daily and Dalton 2003; Rosener 1995). Thus, women are more cooperative and highly motivated to give narrative feedback. On the other hand, there exists other arguments that gender does not typically affect evaluations attitude (Peters et al 1984; Pulakos and Wexley 1983). Therefore, I do not have a signed prediction on the effect of woman evaluators on the amount of narrative comments and suggestion comments.

V. EMPIRICAL RESULTS

5.1. Descriptive Statistics

Figure 1, Panel A shows the time trends of average amounts of narrative comments. While the number of clarification comments show stable pattern over time, both the total narrative comments and suggestion comments monotonically increase over time. I can infer that the primary driver of increasing pattern of total narrative comments might be the pattern of suggestion comments. Panel B shows the amounts of narrative comments according to performance score level. Total amounts of narrative comments do not fluctuate as much as performance score level varies. However, when score is high (low), the amount of suggestion comments increases (decreases), consistent with my predictions on information acquisition cost, confrontation cost, and intrinsic motivation.

[INSERT FIGURE 1 ABOUT HERE]

I present summary statistics for the amount of narrative comments, performance score rate, and characteristics of measurement, evaluator, and firms in Table 2. The average (median) amount of total narrative comments (*#Total*) is 445.7 (417) and its standard deviation is about 210.5. Also, the average (median) amount of suggestion comments (*#Suggest*) is 263.9 (230) and their relative ratio (*Ratio*) is 0.574, on average.

Subjective performance score rate (*Attain*) has the mean value of 0.66, implying lenient rating based on the left skewed distribution (skewness = -0.487). The tenure of evaluators is, on average, about two years. The average fraction of woman evaluators is 8.1 percent, indicating that most evaluators consist of male evaluators. Table 3 presents the Pearson correlation matrix. In general, the determinant variables are not highly correlated with each other.

[INSERT TABLE 2 ABOUT HERE]

[INSERT TABLE 3 ABOUT HERE]

5.2. Determinants of Amounts of Narrative Feedback

Columns (1) to (4) of Table 4 present the estimation results of Equations (1A) and (1B), where the amount of total narrative comments and suggestion comments are the dependent variables. I replace subjective performance score rate ($Attain_t$) with indicator variables of the highest and the lowest quartile of subjective performance score rate ($High\_Attain_t$ and Low_Attain_t) in column (2) and (4). I find that both measures with higher and lower performance score rate tend to exhibit a larger amount of total narrative comments, and measures with higher (lower) performance score rate tend to exhibit a smaller (larger) amount of suggestion comments. These findings are consistent with my hypothesis. For the measures with low performance score, it is easy to write comments because of their low information acquisition cost, high confrontation cost and high probability for improvement, especially suggestions for improvement. Regarding the positive coefficient of high performance score group in column (2), I infer that the concern for confrontation cost toward supervisors is a more powerful driver to generate the total comments at least for high score group. Moreover, the coefficients of *Weight* and *Size* are significantly positive in all columns. These results are consistent with the confrontation cost hypothesis. Evaluators are more likely to concern for bigger SOEs and highly weighted performance measure and hence, they provide more total comments and suggestion comments. In addition, as a performance measure is used repeatedly, the amounts of both total comments and suggestion comments are larger, which is consistent with the information acquisition cost hypothesis. Also, the amount of narrative comments is significantly smaller for larger sized evaluating committee.[32] A member of large team can shirk and generate smaller amount of comments, which supports intrinsic motivation hypothesis. However, the gender and experience effect on the amount of comments are statistically insignificant.[33]

[INSERT TABLE 4 ABOUT HERE]

To determine the relative amount of suggestion comments to the amount of total narrative comments, I estimate Equation (2) using the determinants used in Table 4. As explained above, suggestion for improvement is beneficial in terms of both knowledge effect and acceptability effect. In the presence of prescriptive information, feedback recipients are more likely to adjust their strategies accordingly and work smarter (Audia and Locke 2003). Thus, suggestions for development provide evaluatees with information about developmental opportunities. Also, suggestions comments provide

[32] It is also useful to test the effect of relative team size on the amount of narratives because relative size of evaluating committee per SOE is a more standardized proxy for evaluating team size. When I use the relative size of the committee which is the number of evaluators within the evaluating committee divided by the number of their appointed SOEs, instead of total size evaluating committee, the significantly negative coefficients remain in both regression results.

[33] To properly test information acquisition cost hypothesis, it might be necessary to control evaluator ability of evaluation. I indirectly control for evaluator effect by incorporating evaluator committee fixed effect. When I replace the variables of evaluator characteristics with evaluator fixed effect, the results regarding performance measures and evaluatee characteristics remain unchanged.

evaluatees with more precise and specific information and hence, acceptability of feedback might increase.

As shown in Table 5, the coefficients on *High_Attain*$_t$ and *Low_Attain*$_t$ continue to be significant at the 1 percent level, suggesting that evaluators tend to provide suggestion for development more (less) for the performance measure with low (high) score. These results are consistent with the motivational concern of evaluators. When evaluators face low score problem, they tend to generate more suggestion comments because they provide more knowledge and restore the unpleasantness about the low score. Also, in terms of information acquisition cost, it is relatively easy to provide suggestions in case of the performance measure with low score than high score case. Moreover, the coefficient of *Measure_Age* remains significantly positive and that of *#Evaluator* is significantly negative comparing to the results of Table 4. Also, *Unique*, *Experience*, and *Woman* show significant coefficients. However, *Weight* and *Size* which are related to the confrontation cost hypothesis lose their significance. These results can be explained that relative fraction between suggestion words and clarification words are more sensitive to information acquisition costs and intrinsic motivation of evaluators, rather than confrontation costs.[34] When evaluators face high information acquisition cost

[34] Prior research on feedback has shown that, in an attempt to make the feedback more acceptable to evaluatees, evaluators often deliver negative feedback in more ambiguous terms (Baron 1993; Fisher 1979). Similarly, evaluators might provide less suggestions relative other comments in order to reduce confrontation from evaluatees. Also, there is a pressure of inflating ratings if evaluators provide large number of suggestions and evaluatees successfully incorporate them. Thus, if evaluators are likely to have concerns about future evaluation rating, they would have incentive to provide less suggestions ambiguously. These additional incentives of evaluators can drive insignificant coefficients of variables related to confrontation costs hypothesis in Table 5.

and have high motivation to give suggestion words, they provide more suggestion comments relative to clarification comments.[35]

In sum, I interpret these results as confirming my three hypothesis – information acquisition costs, confrontation costs, and motivation of evaluators.

[INSERT TABLE 5 ABOUT HERE]

5.3. Additional Analysis

5.3.1. Alternative Narrative Units of Analysis

I check the robustness of my analysis on the determinants of narrative comments by using the alternative narrative units. As explained above, I use the number of words as a narrative unit because of its objectivity and reliability in coding process. In line with the principles of qualitative methodology and content analysis (e.g., Crano and Brewer, 1986), each meaningful and distinguishable performance-related theme found in single narrative comments was considered a separate entity. When I also use the number of paragraph as the alternative narrative unit, I find the similar results that support my hypothesis. In untabulated results, I find the significant coefficients for performance score rate, weight, and age among the performance measure determinants, and find big size SOE and evaluating committee with large number of evaluators are significantly associated the amount of narrative comments and suggestions. Moreover, I find the significant coefficient of performance uniqueness in the regression results for the amount of suggestion.

[35] Our test sample is a mixture of observations with zero and one value of our dependent variable (Ratio) as well as positive values. 67 (133) observations of the sample has zero (one) values for the dependent variable. This might be a major violation of the assumption underlying the OLS estimator that the dependent variable is a continuous normal variable. In this case, a Tobit estimator is the correct function (Greene 2000). Tobit regression results are similar to the results shown in Table 5.

5.3.2. Relation between Narrative Feedback and Lenient rating

Evaluators have their own incentives to inflate ratings for poor performers so that they avoid conflicts with subordinates (Ahn and Hyun 2014; Levy and Williams 2004; Mitchell and O'Reilly 1983). Also, the raters are likely to show larger leniency to poor performers than to high performers because motivation effect is greater for poor performers (Bol 2011; Kane 1994; Woods 2012). Therefore, instead of providing detailed narrative feedback, giving lenient rating can mitigate the confrontation costs toward evaluators. To test whether the relationship between narrative feedback and lenient rating is substitutive or not, I additionally create a proxy for leniency and insert it into the Equations (1A) and (1B). I proxy for the leniency tendency by changes in subjective performance score rate which is based on the assumption that the prior performance score might be perceived as a relevant benchmark to its current performance (e.g., Thorsteinson et al. 2008; Woods 2012). Thus, the deviation of current score from prior score might capture the leniency degree (Ahn and Hyun 2014). In untabulated results, I find statistically insignificant coefficients of change in subjective performance score rate in both regressions.[36] These results show that the relation between narrative feedback and lenient rating is neither substitutive nor complementary.[37]

[36] The coefficient of change in score rate is -0.007 (t-statistics = -0.06) (-0.002 (t-statistics = -0.01)) in the regression that has the dependent variable of total comment amount (*ln(#Total)*) (suggestion comment amount (*ln(#Suggest)*)), respectively.

[37] The results can be alternatively interpreted as a mixture of substitutive and complementary relations. While the relation between lenient rating and narrative feedback might be substitutive with respect to the confrontation cost from evaluatees as stated above, the relation can be complementary in terms of concern for supervisors. In other words, evaluators provide more narratives to justify their lenient rating.

VI. CONCLUSION

The purpose of this paper is to examine the determinants of amount of narrative comments and suggestion comments. Using the performance reports dataset of multiple SOEs for multiple years, I investigate the characteristics of performance measures, evaluators, and evaluatees that affect the amount of narrative comments about the evaluatees' performance. I find that (1) evaluators present large amount of narrative comments about the highly weighted, long–tenured, and both high scored and low scored performance measures, and SOEs with large number of evaluators and large number of employees, (2) these findings are similar to the case of suggestion comments excluding the performance score part, (3) the ratio of narrative suggestions to total narrative comments are associated with performance score, measurement age and uniqueness, and evaluators' experience, gender and evaluator group size. Overall, these evidences are consistent with evaluators giving substantial narrative feedback when evaluators face low information acquisition cost, high confrontation cost and have high motivation for providing comments.

The results of this study have several practical implications. I examine the incentive and reaction of evaluators with regard to writing feedback comments. My study shows that the consideration of the written feedback is important in subjective evaluation, and suggests that the incentives of evaluators comes from both passive and active drivers. While concerns towards information acquisition cost and confrontation cost can be considered as passive incentive, helping subordinate by providing suggestion can be regarded as active incentive. In this paper, my findings empirically support that these two incentives exist at the same time when evaluators generate feedback. Also, the intrinsic motivations of both public sector managers and employees are major premise conditions of effort and performance (e.g., Prendergast 2007; Wright 2001), and are higher than that of private sector managers and employees (Lyons et al. 2006). Given its importance in this sector, my results suggest that evaluation system designer should comprehensively consider the motivational effects of evaluators when they provide feedback comments.

My findings are subject to important caveats. Most importantly, generalizability of this study is limited because this study is based only on public entities in Korea and the quality of the variables employed in this study may not be optimal. Societal norms and values in Korea may also have some influence on my results. However, my sample firms employ a wide range of performance measures (e.g., subjective versus objective, common versus unique) and adopt a performance-based bonus system that is parallel to performance evaluation system in most successful enterprises. Second, my analysis covers neither the consequence of narrative feedback with respect to evaluatees nor other characteristics of narrative feedback such as specificity and valence. To investigate the reaction to the feedback information and other characteristics of narrative feedback will be interesting research topics.

APPENDIX A
Example of Coding Process about Narrative Comments

No.	Narrative Comments	#Comment	#Suggest
1	The SOE has the vision of developing the Busan port into a center of logistics in Northeast Asia and reaching the world best port authority that provides the world best service. The business philosophy of CEO such as knowledge business, global business, and productivity improvement is assumed to be partly incorporated into the firm's vision and management objective.	58	0
2	It can be seen that the firm has the systems of performance management and evaluation/incentive management by using KPIs. For example, executive officers of the firm made a management contract with the CEO in June and were scheduled to be paid the differential performance bonus according to the contract. The performance evaluation based on assigned KPIs was also planned to be conducted in the near future. These systems played important roles to unify all the capabilities of employees.	78	0
3	The CEO tried to share his vision and philosophy with employees, but this effort did not bear enough fruit. Because the communication method of sharing those is normally top-down and one-way approach by committee meeting and morning assembly, it was failed to actively collect extensive opinions from employees and share the collected opinions with all the members of the firm. Because the firm was newly established, the vision of CEO was not effectively pooled among the employees.	77	0
4	The CEO understood the treat of the firm such as emerging China effect, the competition with other local port and set up the specific plan to manage the current project. However, there was no long-term strategy to achieve the vision of the firm. It is necessary to check the detailed review about 'Gadukdo new port' project and 'reconstruction of north port'.	61	61
5	It was a very good performance considering the first year of operation. However, revenue was mostly generated from the rent and port dues and the revenue amount is just about 0.5 percent of initial equity. This shows problem of the firm regarding the profitability and long-term financial stability. Therefore, it is needed to review and build up the long-term financial strategy and the specific plan according to the strategy.	69	69
6	In sum, it is evaluated B0 regarding the Leadership performance measure of Busan Port Authority.	.	.
Total		343	130

The table shows the example of coding process about the leadership measure of Busan Port Authority in 2004. This performance measure shows 343 total words (*#Comment*) and 130 suggestion words (*#Suggest*).

REFERENCES

Ahn, T. S., I. Hwang, and M. Kim. 2010. The impact of performance measure discriminability on rate incentives. *The Accounting Review* 85(2): 389-417.

Ahn, T. S., and J. Hyun. 2014. The Effect of Prior Performance Information on Leniency in Subjective Performance Evaluation. Working paper, Seoul National University.

Antonioni, D. 1996. Designing an effective 360-degree feedback appraisal process. *Organizational Dynamics* 25: 24−38.

Atwater, L., and J. Brett. 2006. Feedback format: Does it influence managers' reactions to feedback? *Journal of Occupational and Organizational Psychology* 79: 517-532.

Audia, P. G., and E. A. Locke. 2003. Benefiting from negative feedback. *Human Resource Management Review* 13: 631-646.

Baker, G. P., R. Gibbons, and K. J. Murphy. 1994. Subjective performance measures in optimal incentive contracts. *The Quarterly Journal of Economics* 109(4): 1125-1156.

Banker, R. D., and S. M. Datar. 1989. Sensitivity, precision, and linear aggregation of signals for performance evaluation. *Journal of Accounting Research* 27(1): 21-39.

Baron, R. A. (1993). Criticism informal negative feedback as a source of perceived unfairness in organizations: Effect, mechanisms, and countermeasures. In R. Cropanzano (Ed.), *Justice in the workplace: Approaching fairness in human resource management*. Mahwah, NJ: Lawrence Erlbaum.

Barrett, R. S. 1966. The influence of the Supervisor's Requirements on Ratings1. *Personnel Psychology* 19: 375-387.

Bartol, K. M., and D. A. Butterfield. 1976. Sex effects in evaluating leaders. *Journal of Applied Psychology* 61: 446-454.

Bernardin, H. J., and P. Villanova. 1986. Performance appraisal. In E. A. Locke (Ed.), *Generalizing from laboratory to field settings*. Lexington, MA: Lexington Books.

Bol, J. 2008. Subjectivity in compensation contracting, *Journal of Accounting Literature* 27: 1-24.

Bol, J. 2011. The determinants and performance effects of managers' performance evaluation biases. *The Accounting Review* 86(5): 1549-1575.

Borman,W. C. 1991. Job behaviour, performance, and effectiveness. In M. D. Dunnette and L. M. Hough (Ed.), *Handbook of Industrial and Organizational Psychology, Vol. 2.* Palo Alto, Ca: Consulting Psychologist Press.

Bretz, R. D., G. T. Milkovich, and W. Read. 1992. The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management* 18(2): 321-352.

Brown, E. M. 1968. Influence of training, method, and relationship on the halo effect. *Journal of Applied Psychology* 52: 195-199.

Brutus, S. 2010. Words versus numbers: A theoretical exploration of giving and receiving narrative comments in performance appraisal. *Human Resource Management Review* 20(2): 144-157.

Brutus, S., M. B. Donia, and S. Ronen. 2013. Can Business Students Learn to Evaluate Better? Evidence From Repeated Exposure to a Peer-Evaluation System. *Academy of Management Learning and Education* 12: 18-31.

Burgess, S., and M. Ratto. 2003. The role of incentives in the public sector: Issues and evidence. *Oxford Review of Economic Policy* 19(2): 285-300.

Bushman, R. M., R. J. Indjejikian, and A. Smith. 1996. CEO compensation: The role of individual performance evaluation. *Journal of Accounting and Economics* 21(2): 161-193.

Butler, R. 1987. Task-involving and ego-evolving properties of evaluations: Effects of different feedback conditions on motivated perceptions, interest, and performance. *Journal of Educational Psychology* 79: 474−482.

Cannon, M. D., and R. Witherspoon. 2005. Actionable feedback: Unlocking the power of learning and performance improvement. *The Academy of Management Executive* 19(2): 120-134.

Chomsky, N. 1986. *Knowledge of language: Its nature, origin, and use*. New York, NY: Praeger.

Crano, W. D., and M. B. Brewer. 1986. *Principles and methods of social research*. Boston: Allyn & Bacon.

Daily, C. M., and D. R. Dalton. 2003. Women in the boardroom: a business imperative. *Journal of Business Strategy* 24(5): 8–9.

Dalessio, A. T. (1998). Using multi-source feedback for employee development and personnel decisions. In J. W. Smither (Ed.), *Performance appraisal: State of the art in practice* (pp. 278−330). San Francisco, CA: Jossey-Bass.

Du, F., G. Tang, and S. M. Young. 2012. Influence activities and favoritism in subjective performance evaluation: Evidence from Chinese state-owned enterprises. *The Accounting Review* 87(5): 1555-1588.

Emerson, R. M. 1976. Social exchange theory. *Annual Review of Sociology* 2: 335-362.

Ferris, G. R., and T. A. Judge. 1991. Personnel/human resources management: A political influence perspective. *Journal of Management* 17: 447-488.

Ferstl, K. L., and K. T. Bruskiewicz. 2000. Self-other agreement and cognitive reactions to multirater feedback. *Paper presented at 15th Annual Conference of the Society of Industrial and Organizational Psychology, New Orleans, LA.*

Fisher, C. D. 1979. Transmission of positive and negative feedback to subordinates: A laboratory investigation. *Journal of Applied Psychology* 64: 533-540.

Fitzer, K. 2003. Review of theoretical and applied issues in written language expression. *Canadian Journal of School Psychology* 18: 203−221.

Fletcher, C. R. 1986. Strategies for the allocation of short-term memory during comprehension. *Journal of Memory and Language* 25: 43–58.

Flower, L. S., and J. R. Hayes. 1981. A Cognitive Process Theory of Writing. *College Composition and Communication* 32: 365-387.

Fried, Y., S. Ariel, A. S. Levi, H. A. Ben-David, and R. B. Tiegs. 1999. Inflation of subordinates' performance ratings: main and interactive effects of rater negative affectivity, documentation of work behavior, and appraisal visibility. *Journal of Organizational Behavior* 20(4): 431-444.

Ghorpade, J. 2000. Managing five paradoxes of 360-degree feedback. *The Academy of Management Executive* 14: 140-150.

Gibbs, M., K. A. Merchant, W. A. Van der Stede, and M. E. Vargus. 2004. Determinants and effects of subjectivity in incentives. *The Accounting Review* 79(2): 409-436.

Gilliland, S. W., and J. C. Langdon. 1998. Creating performance management systems that promote perceptions of fairness. In J. W. Smither (Ed.), *Performance appraisal: State of the art and practice* (pp. 209−243). San Francisco, CA: Jossey-Bass.

Gillipsie, T., D. S. Rose, and G. N. Robinson. 2006. Narrative comments in 360 degree feedback: Who says what? *Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.*

Gow, I. D., G. Ormazabal, and D. J. Taylor. 2010. Correcting for cross-sectional and time-series dependence in accounting research. *The Accounting Review* 85(2): 483-512.

Harackiewicz, J. M., and J. R. Larson. 1986. Managing motivation: The impact of supervisor feedback on subordinate task interest. *Journal of Personality and Social Psychology* 51(3): 547.

Harrington, J. E. 2012. Corporate Leniency with Private Information: An Exploratory Example. In Harrington, J. E., and Y. Katsoulakos (Ed.), *Recent Advances in the Analysis of Competition Policy and Regulation*. Cheltenham, Gloucestershire, England: Edward Elgar Publishing.

Harris, M. M. 1994. Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management* 20(4): 737-756.

Holmström, B., and P. Milgrom. 1991. Multi-task principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization* 7: 24-52.

Huber, V. L. 1989. Comparison of the effects of specific and general performance standards on performance appraisal decisions. *Decision Sciences* 20: 545-557.

Ilgen, D., and C. Davis. 2000. Bearing bad news: Reactions to negative performance feedback. *Applied Psychology: An international Review* 49: 550–565.

Kane, J. S. 1994. A model of volitional rating behavior. *Human Resource Management Review* 4(3): 283-310.

Kanfer, R. 1990. Motivation theory and industrial/organizational psychology. In M. D. Dunnette and L. Hough (Ed.), *Handbook of industrial and organizational psychology* (pp. 75−170). Palo Alto, CA: Consulting Psychologists Press.

Keeping, L. M., and P. E. Levy. 2000. Performance appraisal reactions: Measurement, modeling, and method bias. *Journal of Applied Psychology* 85: 708-772.

Kellog, R. T. (1994). *The psychology of writing.* New York, NY: Oxford University Press.

Landy, F. J., and J. L. Farr. 1980. Performance rating. *Psychological Bulletin* 87(1): 72-107.

Larson, J. R. 1984. The performance feedback process: A preliminary model. *Organizational Behavior and Human Performance* 33: 42−76.

Levy, P. E., and J. R. Williams. 2004. The social context of performance appraisal: A review and framework for the future. *Journal of Management* 30(6): 881-905.

Li, F. 2010. The Information Content of forward-looking statements in corporate filings – A naïve Bayesian machine learning approach. *Journal of Accounting Research* 48: 1049-1102.

Li, F. 2011. Textual analysis of corporate disclosures: A survey of the literature. *Journal of Accounting Literature* 29, 143-165.

Lipe, M. G., and S. E. Salterio. 2000. The balanced scorecard: Judgmental effects of common and unique performance measures. *The Accounting Review* 75(3): 283-298.

Longenecker, C., H. Sims, and D. Gioia. 1987. Behind the mask: The politics of employee appraisal. *Academy of Management Executive* 1: 183-193.

Lyons, S.T., Duxbury, L.E. and Higgins, C.A., 2006. A comparison of the values and commitment of private sector, public sector, and parapublic sector employees. *Public Administration Review* 66(4): 605–618.

McShane, S. L., and M. A. Von Glinow. 2008. *Organizational Behavior* (4th Ed). New York: McGraw- Hill/Irwin.

Merchant, K. A., C. Stringer, and P. Theivananthampillai. 2010. Relationships between objective and subjective performance ratings. Working paper, University of Southern California.

Mitchell, T. R., and C. A. O'Reilly III. 1983. Managing poor performance and productivity in organizations. *Research in Personnel and Human Resources Management* (1): 201-234.

Moers, F. 2005. Discretion and bias in performance evaluation: The impact of diversity and subjectivity. *Accounting, Organizations and Society* 30(1): 67-80.

Moxey, L. M., and A. J. Sanford. 2000. Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology* 14: 237−255.

Murphy, K. J. 1999. Executive compensation. *Handbook of Labor Economics* 3: 2485-2563.

Murphy, K. R. and J. N. Cleveland. 1995. *Understanding Performance Appraisal*, Thousand Oaks, CA: Sage Publications.

Nishimura, A., 2003. *Management Accounting; Feed Forward and Asian Perspectives*. Palgrave Macmillan, Great Britain.

Otley, D. 1999. Performance management: a framework for management control systems research. *Management Accounting Research* 10: 363–382.

Peters, L. H., E. J. O'Connor, J. Weekley, A. Pooyan, B. Frank, and B. Erenkrantz. 1984. Sex bias and managerial evaluations: A replication and extension. *Journal of Applied Psychology* 69(2): 349-352.

Pitkänen, H., and K. Lukka. 2011. Three dimensions of formal and informal feedback in management accounting. *Management Accounting Research* 22(2): 125-137.

Prendergast, C., 2007. The motivation and bias of bureaucrats. *American Economic Review* 97(1): 180–196.

Prendergast, C., and R. H. Topel. 1996. Favoritism in organizations. *Journal of Political Economy* 104(5): 958-978.

Pulakos, E. D., and K. N. Wexley. 1983. The relationship among perceptual similarity, sex, and performance ratings in manager-subordinate dyads. *Academy of Management* 26(1): 129-139.

Rogers, W. 1993. Regression standard errors in clustered samples. *Stata Technical Bulletin* 13: 19-23.

Rosener, J. B. 1995. *America's Competitive Secret: Utilizing Women as a Management Strategy*. Oxford University Press: New York.

Sanford, A. J., N. Fay, A. Stewart, and L. Moxley. 2002. Perspectives in statements of quantity, with implications for consumer psychology. *Psychological Science* 13: 130−134.

Schmitt, N., R. J. Klimoski, G. R. Ferris, and K. M. Rowland. 1991. *Research methods in human resources management*. South-Western Publishing Company.

Schneier, C. E. 1977. Operational utility and psychometric characteristics of BES: A cognitive reinterpretation. *Journal of Applied Psychology* 62(5): 541-548.

Sedatole, K., and A. Woods. 2013. Consistency and Organizational Justice: The role of calibration committees in subjective performance evaluation systems. Working paper, Michigan State University.

Simon, H. 1997. *Administrative Behavior.* 4th Ed. The Free Press. New York, NY.

Smither, J. W. 1998. *Performance appraisal: State of the art in practice.* San Francisco, CA: Jossey-Bass.

Smither, J. W., M. London, and R. R. Reilly. 2005. Does performance improve following multisource feedback? A theoretical model, meta-analysis and review of empirical findings. *Personnel Psychology* 58: 33-66.

Smither, J. W., and A. G. Walker. 2004. Are the characteristics of narrative comments related to improvement in multirater feedback over time? *Journal of Applied Psychology* 89: 575-581.

Spence, J. R., and L. M. Keeping. 2011. Conscious rating distortion in performance appraisal: A review, commentary, and proposed framework for research. *Human Resource Management Review* 21: 85-95.

Stetz, T. A., and J. M. Ford. 2010. Leadership and same-gender bias: A content analysis of promotion recommendations. *Journal of Psychological Issues in Organizational Culture* 1(3): 6-18.

Thorsteinson, T. J., J. Breier, A. Atwell, C. Hamilton, and M. Privette. 2008. Anchoring effects on performance judgments. *Organizational Behavior and Human Decision Processes* 107: 29-40.

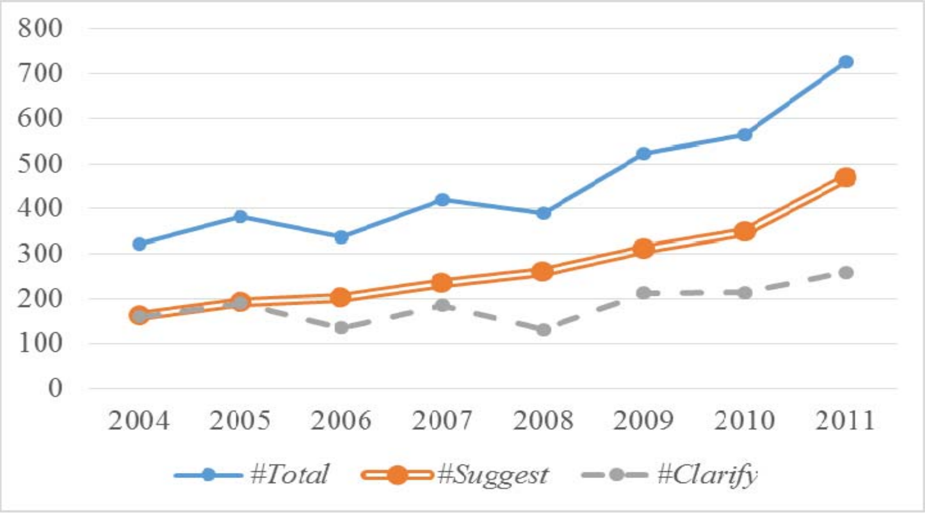
Toegel, G., and J. A. Conger. 2003. 360-degree assessment: Time for reinvention. *Academy of Management Learning and Education* 2(3): 297-311.

Woods, A. 2012. Subjective adjustments to objective performance measures: The influence of prior performance. *Accounting, Organizations and Society* 37: 403-425.

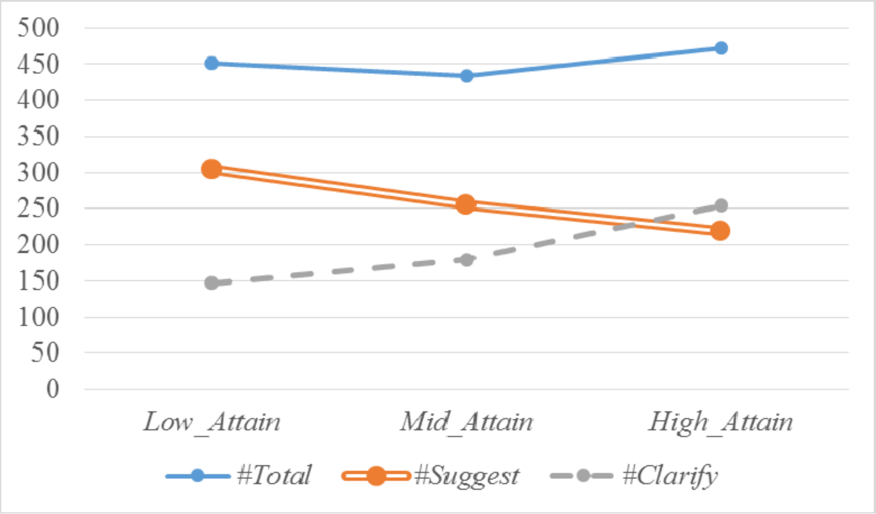
Wright, B.E., 2001. Public-sector work motivation: a review of the current literature and a revised conceptual model. *Journal of Public Administration Research and Theory* 11(4): 559–586.

FIGURE 1
Amounts of Narrative Feedback according to Time and Score Ratio

Panel A: Time Trends of Average Amounts of Narrative Feedback



Panel B: Average Amounts of Narrative Feedback according to Score Ratio (*Attain*)



The sample includes 2,120 subjective performance measures (*Attain*), number of total words (*#Total*), number of suggestion words (*#Suggest*), and number of clarification words (*#Clarify*) from 2004 to 2011 in both Panel A and Panel B.

TABLE 1
Examples of Performance Rating and Composition of Evaluators

Panel A: Example of Performance Ratings for an SOE in 2005[a]

Performance Category	Individual Measure	Measurement[b]	Commonality	Weight	Rating	Score	Attain (=Score rate)
1. Overall management	Labor productivity	OBJ	Common	3.5	-	0.000	0.000
	Customer satisfaction (OBJ)	OBJ	Common	4.2	-	3.394	0.808
	Customer satisfaction (SUB)	SUB	Common	6.3	A0	5.513	0.875
	Managing board of directors	SUB	Common	7	C	3.500	0.500
	…						
2. Main business	Maintenance of park facility	OBJ	Unique	5.5	-	5.500	1.000
	Effort for improvement of park environment	SUB	Unique	2.75	B+	2.063	0.750
	Effort for protection of nature	SUB	Unique	3.15	B0	1.969	0.625
	…						
3. Business management	Financial structure	OBJ	Common	3.3	-	2.364	0.716
	Labor union management	SUB	Common	6.5	B0	4.063	0.625
	Budget management	SUB	Common	3.95	B0	2.469	0.625
	…						
SUB/C total				54.95		33.171	
SUB/U total				10.05		7.138	
OBJ/C total				13.55		7.304	
OBJ/U total				21.45		21.222	
Total				100		68.835	

Panel B: Example of Composition of Evaluators in 2008[c]

SOE Class	Subjective measure category			Objective measure	No. of evaluators
	Overall management	Main business	Business management		
Overall	3 professors				3
1. SOC (14 SOEs)	4 professors 2 industrial experts	5 professors 2 industrial experts	6 professors	2 professors 2 CPAs	23
2. Service (10 SOEs)	4 professors 1 industrial expert	4 professors 1 industrial expert	5 professors	1 professor 3 CPAs	19
3. Inspection management (6 SOEs)	3 professors	4 professors 1 industrial expert	3 professors	2 CPAs	13
4. Culture (9 SOEs)	3 professors 1 industrial expert	4 professors	3 professors 1 industrial expert	1 professor 2 CPAs	15
5. Industry promotion I (10 SOEs)	6 professors	4 professors	3 professors 1 CPA	3 CPAs	17
6. Industry promotion II (6 SOEs)	3 professors	3 professors	2 professors 1 industrial expert	1 professor 1 CPA	11
7. Education (6 SOEs)	2 professors 1 industrial expert	3 professors	2 professors 1 industrial expert	1 professor 2 CPAs	12

8. Pension (14 SOEs)	5 professors 1 industrial expert	5 professors 1 industrial expert	7 professors	4 CPAs	23
Total (75 SOEs)	33 professors 6 industrial experts	32 professors 5 industrial experts	31 professors 3 industrial experts 1 CPA	6 professors 19 CPAs	136

Panel A shows the example of performance ratings for Korea National Park Service in 2005. OBJ and SUB in the measurement column in Panel A are objective and subjective performance measures, respectively. Panel B presents the example of evaluators composition in 2008.

TABLE 2
Descriptive Statistics of the Sample

Variables [a]	N	Mean	Standard Deviation	Min	Q1	Median	Q3	Max
#Total	2,120	445.700	210.512	60	294	417	570	1,550
#Suggest	2,120	263.941	180.332	0	130	230	359	1,244
#Clarify	2,120	181.759	135.409	0	81	156	255	929
Ratio	2,120	0.574	0.252	0.000	0.393	0.578	0.773	1.000
Attain	2,120	0.660	0.127	0.000	0.600	0.650	0.750	1.000
Unique	2,120	0.305	0.460	0.000	0.000	0.000	1.000	1.000
Weight	2,120	3.730	1.763	0.420	3.000	3.000	4.800	12.000
Measure_Age	2,120	3.476	2.127	1.000	2.000	3.000	5.000	8.000
Size	2,120	7.066	0.998	4.804	6.623	7.019	7.466	9.426
Experience	2,120	2.051	0.399	1.259	1.800	2.053	2.231	3.063
Expert	2,120	0.193	0.084	0.000	0.146	0.163	0.250	0.400
#Evaluator	2,120	27.233	18.352	10.000	13.000	16.000	43.000	71.000
Woman	2,120	0.081	0.052	0.000	0.067	0.077	0.125	0.167

The sample includes 2,120 unique measure-firm-years and 17 unique SOEs from 2004 to 2011. Data for narrative feedback, performance rating and evaluator characteristics are manually collected from the performance evaluation report of SOEs.

a Variable Definitions:

#Total = number of total words of the subjective measure;

#Suggest = number of suggestion words of the subjective measure;

#Clarify = number of clarification words of the subjective measure, which is the number of non-suggestion words (= number of total words – number of suggestion words);

Ratio = fraction of suggestion words, which is the number of suggestion words divided by the number of total words of the subjective measure;

Attain = performance score rate of the subjective measure;

Unique = 1 if the measure is a unique one, 0 if it is a common one;

Weight = weight assigned to the measure;

Measure_Age = tenure of a subjective measure, which is the number of years each performance measure has been with the SOE;

Size = natural logarithm of the number of employees;

Experience = average tenure of evaluators within the evaluating committee;

#Evaluator = number of evaluators within the evaluating committee;

Expert = Fraction of industrial experts within the evaluating committee, which is the number of industrial experts divided by the total number of evaluators within the evaluating committee; and

Woman = fraction of female evaluators, which is the number of female evaluators within an SOE class divided by the number of total evaluators within the SOE class.

TABLE 3
Pearson Correlation Matrix

	(1) #Total	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*	*(9)*	*(10)*	*(11)*	*(12)*
(2) #Suggest	0.77***											
	(0.00)											
(3) #Clarify	0.53***	-0.13***										
	(0.00)	(0.00)										
(4) Ratio	0.16***	0.68***	-0.66***									
	(0.00)	(0.00)	(0.00)									
(5) Attain	0.21***	0.01	0.30***	-0.22***								
	(0.00)	(0.58)	(0.00)	(0.00)								
(6) Unique	-0.03	-0.08***	0.06***	-0.12***	0.01							
	(0.18)	(0.00)	(0.01)	(0.00)	(0.52)							
(7) Weight	0.42***	0.29***	0.27***	0.01	0.09***	-0.20***						
	(0.00)	(0.00)	(0.00)	(0.52)	(0.00)	(0.00)						
(8) Measure_Age	0.43***	0.40***	0.13***	0.18***	0.21***	-0.11***	0.05**					
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)					
(9) Size	-0.04**	-0.03	-0.02	-0.01	0.10***	0.05**	-0.01	0.01				
	(0.04)	(0.11)	(0.29)	(0.49)	(0.00)	(0.02)	(0.60)	(0.58)				
(10) Experience	0.11***	0.14***	-0.02	0.16***	0.10***	0.01	-0.04**	0.30***	0.05**			
	(0.00)	(0.00)	(0.39)	(0.00)	(0.00)	(0.72)	(0.05)	(0.00)	(0.04)			
(11) Expert	-0.06***	-0.08***	0.01	-0.06***	-0.01	-0.05**	0.07***	-0.20***	0.08***	-0.36***		
	(0.01)	(0.00)	(0.65)	(0.00)	(0.81)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)		
(12) #Evaluator	-0.24***	-0.27***	-0.02	-0.19***	0.02	-0.02	0.04*	-0.32***	0.03	-0.41***	0.33***	
	(0.00)	(0.00)	(0.26)	(0.00)	(0.46)	(0.32)	(0.06)	(0.00)	(0.11)	(0.00)	(0.00)	
(13) Woman	0.10***	0.13***	-0.03	0.08***	0.04*	0.03	0.00	0.11***	0.26***	-0.21***	0.33***	-0.07***
	(0.00)	(0.00)	(0.23)	(0.00)	(0.09)	(0.22)	(0.92)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

P-values are reported in parentheses under each estimated correlation value. The symbols *, **, and *** correspond to 10 percent, 5 percent, and 1 percent significance levels, respectively, for two-tailed t-tests.

TABLE 4
Regressions Estimating the Determinants of Narrative Comments Amounts [a]

	Dependent Variable					
Independent Variables[b]	Predicted	(1)	(2)	Predicted	(3)	(4)
	Sign	*ln(#Total+1)*		Sign	*ln(#Suggest+1)*	
Attain	?	0.489***		–	-0.844***	
		(7.50)			(-6.30)	
Low_Attain	H1H2H3(+)		0.022***	H1H2H3(+)		0.185***
			(3.02)			(8.45)
High_Attatin	H1H3(–) H2(?)		0.078**	H1H2H3(–)		-0.367***
			(2.55)			(-3.80)
Unique	H1H2(–)	-0.001	0.004	H1H2(–)	-0.105	-0.103
		(-0.02)	(0.08)		(-1.63)	(-1.60)
Weight	H2(+)	0.105***	0.108***	H2(+)	0.125***	0.133***
		(11.26)	(10.70)		(8.57)	(11.11)
Measure_Age	H1(+)	0.077***	0.080***	H1(+)	0.127***	0.118***
		(4.60)	(4.44)		(3.75)	(3.68)
Size	H2(+)	0.388**	0.390**	H2(+)	0.582**	0.502**
		(2.38)	(2.39)		(2.19)	(2.15)
Experience	H1(+)	-0.037	-0.015	H1(+)	0.264	0.215
		(-0.31)	(-0.12)		(1.30)	(1.13)
Expert	H1(+)	0.482	0.441	H1(+)	0.473	0.648
		(0.98)	(0.95)		(0.91)	(1.15)
#Evaluator	H3(–)	-0.007***	-0.008***	H3(–)	-0.014***	-0.013***
		(-2.98)	(-2.88)		(-2.88)	(-2.92)
Woman	H3(?)	0.362	0.420	H3(?)	1.423	1.227
		(0.44)	(0.52)		(1.16)	(0.97)
Intercept		1.936	2.118*		-0.171	-0.037
		(1.54)	(1.69)		(-0.07)	(-0.02)
Fixed Effect		SOE	SOE		SOE	SOE
Number of obs		2,120	2,120		2,120	2,120
Adjusted R^2		0.407	0.397		0.207	0.221

Z-statistics are reported in parentheses under each estimated coefficient. Standard errors are corrected for heteroskedasticity using the Huber-White robust standard errors clustered by year and performance measure. The symbols *, **, and *** correspond to 10 percent, 5 percent, and 1 percent significance levels, respectively, for two-tailed tests. Please refer to the paper for a detailed explanation of these tests.

a This table shows the coefficient estimates of the leniency determinants by using the following equations:

$$ln(\#Total{+}1)_{ijt} = \alpha_0 + \alpha_1 Attain_{ijt} + \alpha_2 Unique_{ijt} + \alpha_3 Weight_{ijt} + \alpha_4 Measure\_Age_{ijt} + \alpha_5 Size_{jt} + \alpha_6 Experience_{jt} + \alpha_7 \#Evaluator_{jt} + \alpha_8 Expert_{jt} + \alpha_9 Woman_{jt} + Fixed\ Effects + \varepsilon_{ijt}. \quad (1A)$$

$$ln(\#Suggest{+}1)_{ijt} = \alpha_0 + \alpha_1 Attain_{ijt} + \alpha_2 Unique_{ijt} + \alpha_3 Weight_{ijt} + \alpha_4 Measure\_Age_{ijt} + \alpha_5 Size_{jt} + \alpha_6 Experience_{jt} + \alpha_7 \#Evaluator_{jt} + \alpha_8 Expert_{jt} + \alpha_9 Woman_{jt} + Fixed\ Effects + \varepsilon_{ijt}. \quad (1B)$$

where *i*, *j*, and *t* indicate performance measure *i*, firm *j*, and year *t*, respectively.

b The independent variable, Low_Attain_t ($High\_Attain_t$) is an indicator variable which refers to 1 if the score rate belongs to the lowest (highest) quartile of *SUB_Attain* in the current year, 0 otherwise. See Table 2 for the definitions of other variables.

TABLE 5
Regressions Estimating the Determinants of Narrative Comments Ratio [a]

Independent Variables [b]	Predicted Sign	Dependent Variable
		Ratio *(= #Suggest / #Total)*
Low_Attain	H1H2H3(+)	0.062***
		(5.29)
High_Attatin	H1H2H3(–)	-0.128***
		(-5.77)
Unique	H1H2(–)	-0.060***
		(-3.37)
Weight	H2(+)	0.004
		(1.34)
Measure_Age	H1(+)	0.009**
		(1.98)
Size	H2(+)	0.082
		(0.86)
Experience	H1(+)	0.057***
		(2.64)
Expert	H1(+)	0.046
		(1.00)
#Evaluator	H3(–)	-0.001***
		(-3.28)
Woman	H3(?)	0.296*
		(1.88)
Intercept		-0.169
		(-0.23)
Fixed Effect		SOE
Number of obs		2,120
Adjusted R^2		0.168

Z-statistics are reported in parentheses under each estimated coefficient. Standard errors are corrected for heteroskedasticity using the Huber-White robust standard errors clustered by year and performance measure. The symbols *, **, and *** correspond to 10 percent, 5 percent, and 1 percent significance levels, respectively, for two-tailed tests. Please refer to the paper for a detailed explanation of these tests.

a This table shows the coefficient estimates of the leniency determinants by using the following equations:

$$Ratio_{ijt} = \alpha_0 + \alpha_1 Attain_{ijt} + \alpha_2 Unique_{ijt} + \alpha_3 Weight_{ijt} + \alpha_4 Measure\_Age_{ijt} + \alpha_5 Size_{jt} + \alpha_6 Experience_{jt} + \alpha_7 \#Evaluator_{jt} + \alpha_8 Expert_{jt} + \alpha_9 Woman_{jt} + Fixed\ Effects + \varepsilon_{ijt}. \quad (2)$$

where *i*, *j*, and *t* indicate performance measure *i*, firm *j*, and year *t*, respectively.

b The independent variable, Low_Attain_t ($High\_Attain_t$) is an indicator variable which refers to 1 if the score rate belongs to the lowest (highest) quartile of *SUB_Attain* in the current year, 0 otherwise. See Table 2 for the definitions of other variables.

국문초록

주관적 성과평가에서의 관대화 경향과 서술적 피드백

본 학위논문은 주관적 성과평가(subjective performance evaluation)에서의 평가자 인센티브에 대해 고찰한다. 주관적 성과평가는 평가자가 주관적으로 특정 성과지표에 대한 성과를 척도 점수로 평가하는 것으로, 목표설정을 중심으로 한 객관적 성과평가가 내포하는 문제점을 보완하는 역할을 하고, 그 자체로 중요한 성과지표를 구성하는 한 요소이다. 이처럼 성과평가에 있어서 중요한 주관적 성과평가와 관련해서, 평가자의 인센티브 관점에서 주관적 성과평가 프로세스에 대해 알려진 바가 많지 않다. 주관적 성과평가의 과정은 평가자가 피평가자의 성과 정보를 수집하고 해석하여, 평가 척도 점수를 제공하고 그에 대한 피드백을 제공하는 것이다. 이 과정에서 평가자가 본인의 효용을 고려하여 평가척도를 본래 판단한 성과 점수 보다 관대하게 부여할지 여부를 결정하고, 또한 서술적 피드백을 간략하게 혹은 자세하게 제공하기도 한다. 본 학위논문에서는 우리나라의 공기업 경영실적 평가제도상 공시한 성과평가 데이터를 바탕으로 주관적 성과평가 과정에서 평가자의 인센티브를 분석한다.

첫 번째 연구는 주관적 평가 과정에서 평가 척도 점수를 부여할 때 발생할 수 있는 평가자의 관대화 경향이 전년도 성과 정보에 영향을 받는가에 대해 분석한다. 성과평가와 관련한 선행연구가 객관적 성과지표 점수 또는 척도 점수의 중

간 값을 관대화 경향의 기준점(benchmark)으로 이용하고, 횡단면 연구(cross-sectional research)에 국한되었기 때문에 관대화 경향의 지속성 여부 및 전년도 성과 정보의 당해 년도 관대화 경향에의 영향을 분석하는 데에는 한계가 있었다. 본 학위논문에서는 당해 년도 성과 점수와 상관관계가 높은 전년도 성과 점수를 기본으로 한 기준점을 바탕으로 관대화 경향을 계측한다. 이와 같이 새로 개발된 관대화 경향 변수를 바탕으로 관대화 경향이 시간이 지나도 지속되고, 전년도에 성과 점수가 낮고(높고), 덜(더) 관대하게 평가 받은 피평가자에 대해 당해 년도에는 관대화 경향이 크다(작다)는 실증 분석 결과를 제시한다. 이와 같은 결과는 평가자의 인센티브가 전년도 성과 정보에 따라 좌우될 수 있고, 그 경향이 지속된다는 점을 시사한다.

두 번째 연구는 주관적 성과평가 과정 중 시행되는 서술적 피드백(narrative feedback)의 양에 대한 결정요인을 분석한다. 공기업 경영실적평가제도에 의하면, 평가자가 피평가자인 공기업의 각 주관적 성과지표 별로 척도 점수의 근거 및 개선사항을 서술하도록 하고 있다. 본 연구에서는 이 서술적 피드백 데이터를 바탕으로, 평가자는 본인의 인센티브 관점에서 서술에 필요한 평가 정보의 획득 비용(information acquisition cost)이 작을수록, 피평가자 및 상위 평가 감시단에 대한 대면비용(confrontation cost)이 클수록, 서술적 피드백 제공에 대해 평가자 본인의 내재적 동기(intrinsic motivation)가 높을수록 서술적 피드백 및 개선사항을 많이 제공하는 경향이 있다는 분석 결과를 제시한다.

본 학위논문의 발견은 성과평가를 둘러싼 이해관계자에게 유용한 시사점을 제공한다. 특히 평가 시스템을 설계하는 담당자 입장에서, 직전 년도 성과 정보에 따라 평가자가 체계적인 편향(bias)을 보이고, 그 편향이 지속될 수 있다는 점, 그리고 또한 평가자의 인센티브에 따라 서술적 피드백의 양이 변동될 수 있음을 시스템 설계 및 관리 시 감안할 필요가 있을 것이다. 피평가자의 성과향상에 도움이 되는 관대화 경향, 서술적 피드백의 활용 방법의 개발은 후속 연구주제로 남겨둔다.

주요어: 성과평가, 주관적 성과지표, 과거성과, 관대화 경향, 서술적 피드백, 개선사항, 정보획득비용, 대면비용.

학번: 2010-30157