Available online at www.sciencedirect.com
SCIENCE DIRECT®
Accounting, Organizations and Society 30 (2005) 67–80
Accounting, Organizations and Society
www.elsevier.com/locate/aos

Discretion and bias in performance evaluation: the impact of diversity and subjectivity

Frank Moers

Faculty of Economics and Business Administration, MARC/Department of Accounting & Information Management, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands

Abstract

This paper examines the impact of performance measure diversity and the use of subjective performance measures on performance evaluation bias. The empirical results indicate that both performance measure diversity and subjectivity are positively related to performance evaluation bias. More specifically, I find that the use of multiple objective performance measures and the use of subjective performance measures are related to more compressed performance ratings and more lenient performance ratings, which could result in problems in personnel decisions and future incentives.

Introduction

The design of incentive systems has attracted considerable attention in accounting research. The focus of this research is on the choice of performance measures in incentive systems (e.g., Govindarajan, 1984; Ittner, Larcker, & Rajan, 1997) and the effects that this choice has (e.g., Govindarajan & Gupta, 1985; Wallace, 1997). Agency theory states that any (costless) performance measure that is informative about the agent's effort should be used for incentive purposes (Holmström, 1979). Because no single performance measure is likely to be complete, the informativeness principle states that incentive contracts should include multiple performance measures. Especially the discussion about the incompleteness of financial performance measures has led to the call for the use of multiple performance measures (e.g., Kaplan & Norton, 1996).

Despite this increased call for the use of multiple performance measures, there is limited empirical evidence of how these measures are combined in an overall evaluation and subsequently tied to compensation. Ittner and Larcker (1998) indicate that 38% of the respondents to a Towers Perrin survey state that they experience problems in weighting the different performance measures in a balanced scorecard. Further, Ittner, Larcker, and Meyer (2003) and Lipe and Salterio (2000) show that when multiple performance measures are used and superiors have discretion in weighting the different performance measures, there is a general tendency to overemphasize objective and common measures of performance. This paper extends the incentive literature by examining the impact of performance measure diversity and the use of subjective performance measures on performance evaluation bias. Performance measure diversity is

E-mail address: f.moers@aim.unimaas.nl (F. Moers).

0361-3682/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aos.2003.11.001

defined in this paper as the use of multiple performance measures for incentive purposes. Subjective performance measures, on the other hand, are defined as superior's subjective judgments about qualitative performance indicators.

Analytical studies indicate that incentive contracting can be improved by incorporating more diverse performance measures, including the use of subjective performance measures. For example, Feltham and Xie (1994) and Datar, Cohen Kulp, and Lambert (2001) show that the use of alternative performance measures can improve the effort allocation of the agent, which leads to more goal congruence. Baker, Gibbons, and Murphy (1994) indicate that the combined use of objective performance measures (explicit contracts) and subjective performance measures (implicit contracts) is, in some circumstances, optimal. Similarly, Baiman and Rajan (1995) show that the use of both contractible (objective) and non-contractible (subjective) information in bonus pool arrangements leads to Pareto improvements compared to a situation where only contractible information is used.

Although increasing the number of performance measures and using subjective performance measures can provide more efficient incentives, it also provides the principal with more discretion in performance evaluation. Most analytical studies assume that the principal will not renege on promised incentive payments when having discretion either because of reputation concerns (Baker et al., 1994) or because he commits to a fixed bonus pool (Baiman & Rajan, 1995). However, superiors in organizations often are not the residual claimants of subordinates' outputs and therefore have no incentives to renege but rather have incentives to bias the performance evaluation (Prendergast & Topel, 1993). Research in psychology indicates that superiors compress performance ratings and give more lenient performance ratings when these are used for incentive purposes (Jawahar & Williams, 1997; Milkovich & Newman, 1993). This bias in performance evaluation is problematic because it becomes more difficult to make the right personnel decisions, such as promotions. The problem of bias has only relatively recently been addressed in the economics literature (e.g., Prendergast & Topel, 1993, 1996) and there is virtually no empirical evidence of how performance measurement affects bias. Given the increased call for more diverse performance measures, including the use of subjective performance measures, it seems warranted to examine how performance measure diversity and subjectivity affect performance evaluation bias.

I use a proprietary archival data set of one firm to examine the impact of performance measure diversity and subjectivity on performance evaluation bias. The data set used is unique in the sense that it provides detailed information about a number of components of the subordinate-specific incentive contract, such as, the 'incentive weights' for both objective and subjective performance measures and the number of objective and subjective performance measures. The empirical results indicate that both performance measure diversity and subjectivity are positively related to performance evaluation bias. More specifically, I find that the use of multiple objective performance measures and the use of subjective performance measures are related to more compressed performance ratings and more lenient performance ratings. These results suggest that increasing the number of performance measures and using subjectivity in performance evaluation lead to evaluations that make it more difficult to differentiate among subordinates, which could result in problems in personnel decisions and future incentives.

This paper contributes to the literature in several ways. First, it is one of the few studies that examines the effects performance measure diversity. Ittner and Larcker (1999) examine the effects of performance measure diversity on incentive plan outcomes, such as perceived financial benefits and plan terminations. They find little support for the benefits of using multiple (non-financial) performance measures. This paper extends their study by examining *why* this might be the case. Second, this paper addresses a topic that has recently come to the attention of the economics literature on incentive systems, i.e., discretion and bias, and it extends the theoretical literature by empirically examining how characteristics of performance measurement systems affect bias in performance evaluation. Finally, it extends a small but growing

stream of empirical research on the use of subjectivity in incentive contracts (e.g., Gibbs, Merchant, Van der Stede, & Vargus, 2002; Hayes & Schaefer, 2000; Murphy & Oyer, 2001). These studies predominately focus on the determinants of the use of subjectivity, while the current study focuses on the effects.

The remainder of this paper is organized as follows. In Theory and Hypothesis Development section, I describe the theoretical background and develop hypotheses. In Research Method section, I discuss the research site, incentive plan, and the data collection. In Results section, I present the empirical results, and finally in Discussion and Conclusion section, I discuss the implications of the empirical results for scorecard-type of performance measurement systems and provide a conclusion.

Theory and hypothesis development

Diversity, subjectivity and incentives

The dominant goal of incentive systems is to give employees incentives to provide effort and to subsequently differentiate among the highly skilled and less skilled employees. Agency theory predicts that by relating pay to performance, employees are motivated to exert more effort in order to increase pay through improved performance (Holmström, 1979). The incentive effects of incentive contracts are determined by the performance measures used because employees direct their attention to those aspects of the job that are being measured. As a result, the choice of performance measures is crucial in providing the correct incentives. The informativeness principle indicates that any performance measure that provides (incremental) information about the employee's actions should be used for incentive purposes. Since no single performance measure is likely to be complete (Baker et al., 1994; Kaplan & Norton, 1996), the informativeness principle predicts that incentive contracting is improved by incorporating a more diverse set of performance measures. For example, objective performance measures, like financial performance measures, are only informative about the measurable aspects of an employee's job and provide no incentives for the more qualitative aspects, like cooperation and innovation. Subjective performance measures, on the other hand, are informative about the qualitative job aspects and are therefore of value to incentive contracts since they provide incentives not provided by objective performance measures (Baiman & Rajan, 1995; Baker et al., 1994). If objective and subjective performance measures are informative about different aspects of the agent's job, then the use of both measures in incentive contracts leads to greater effort intensity.

Furthermore, performance measure diversity can improve incentive contracting by alleviating the problem of goal congruence (Datar et al., 2001; Feltham & Xie, 1994). Feltham and Xie (1994) illustrate that additional performance measures can affect the agent's effort allocation to make it more congruent with the principal's objectives, which makes these measures valuable for incentive purposes. In addition, Datar et al. (2001) show that the use of multiple performance measures allows the principal to choose incentive weights that maximize the congruence between the agent's compensation and the firm's outcome. Thus, diversity in performance measurement can lead to both greater effort intensity and a reduction in non-congruity by affecting the effort allocation.

Diversity, subjectivity and performance evaluation bias

The previous discussion of the value of diversity and subjectivity in incentive contracts is based on agency theory, which traditionally assumes that an honest principal contracts with an agent who cannot be trusted. The assumption of an honest principle becomes crucial if the principal has discretion in performance evaluation and incentive contracts are implicit (Prendergast & Topel, 1993). In general, both subjective performance measures and diversity in performance measurement provide superiors with discretion in performance evaluation. Subjective performance measures provide the superior with discretion because no clear performance standards exist for these measures and assessed performance is solely determined by

subjective judgments. Similarly, more diversity in performance measurement gives the superior a portfolio of performance measures that is likely to consist of partly conflicting outcomes. As a result, the superior has the opportunity to ex post attach different weights of importance to each measure and give a performance rating that he sees fit. If the superior can commit to being honest when subjectively assessing the agent's performance, then any implicit contract is in fact explicit. However, this assumption appears to be inconsistent with empirical evidence, which indicates that discretion in performance evaluation gives rise to a number of problems (Prendergast & Topel, 1993).

Although the most obvious problem seems to be reneging, which means that contracted performance is not rewarded, the incentives to renege are often non-existent for superiors because they are not the residual claimants of subordinates' output (Prendergast & Topel, 1993). A more important problem is the issue of performance evaluation bias. The mere fact that superiors are not the residual claimants implies that superiors have an opportunity to let their preferences determine the allocation of rewards. Previous research in psychology suggests that performance ratings are lenient when these ratings are used for administrative purposes such as incentive pay and promotion decisions (Jawahar & Williams, 1997). Furthermore, superiors often insufficiently differentiate among subordinates, leading to a compression of performance ratings (Milkovich & Newman, 1993). The incentives of superiors to bias the performance evaluation of subordinates often relate to the psychological cost of communicating poor performance, favoritism, and preferences for equity in rewards (Prendergast & Topel, 1993).

Bias in performance evaluation is problematic because there are not only direct costs associated with bias but also indirect costs. The direct costs relate to higher compensation costs than those warranted by the 'true' performance of the subordinates. The indirect costs relate to the difficulty of making important personnel decisions based on the performance ratings and the impact of incentives on motivation. If the performance ratings are biased, then 'all' employees seem to be above average performers and it becomes difficult to select the 'right' subordinate for the 'right' job. Further, if subordinates become aware of the bias, they might become less motivated and therefore provide less effort in the future. Since personnel decisions and incentives are important determinants of firm performance, the indirect costs of bias can be substantial and are likely to be much higher than the direct costs. Bias in performance evaluation is therefore an important aspect to consider when designing incentive contracts.

Hypotheses

Although previous evidence suggests that bias in performance ratings exists when these ratings are used for administrative purposes, there is limited evidence of how performance measurement affects bias. Since diversity in performance measurement and subjective performance measures give the superior discretion in performance evaluation, it seems warranted to examine if diversity and subjectivity affect performance evaluation bias. In this paper, I specifically focus on performance evaluation bias in terms of leniency and compression of performance ratings. Based on the previous discussion, I expect that the use of subjective performance measures per se (subjectivity) leads to more lenient and more compressed performance ratings. I further expect that the use of multiple *objective* performance measures (diversity) leads to more lenient and more compressed performance ratings. I do not expect that the number of *subjective* performance measures affects bias because subjectivity per se already provides the superior with discretion. Although this latter expectation is not explicitly hypothesized, it is part of the empirical tests. As a result, the following hypotheses apply. [1]

[1] Although this study is not a test of the Baker et al. (1994) model, it is related to their distinction between subjective weights on objective performance measures and objective weights on subjective performance measures. More specifically, hypotheses 1A and 2A relate to the effects of subjective weights on objective performance measures, while hypotheses 1B and 2B relate to the effects of objective weights on subjective performance measures.

H1A: *Performance measure diversity, with respect to objective performance measures, leads to more lenient performance ratings.*

H1B: *Subjectivity in performance measurement leads to more lenient performance ratings.*

H2A: *Performance measure diversity, with respect to objective performance measures, leads to more compressed performance ratings.*

H2B: *Subjectivity in performance measurement leads to more compressed performance ratings.*

Research method

Research site

The research site used in this study is a privately held Dutch industrial firm focused on maritime activities. The firm, hereafter called MARITCORP, was founded in 1875 and has since then always been active in shipbuilding and ship conversions. MARITCORP is primarily located in The Netherlands, although it has also subsidiaries in Belgium and the United Kingdom. [2] MARITCORP's major activities are focused on shipbuilding, maintenance and repair of ships, construction of oil and gas extraction installations for offshore and onshore fields, machining of intermediate and finished parts for ships, design and manufacturing of high-grade gear transmissions for ships, and technical services in the field of materials and welding technology. The peripheral activities of MARITCORP include the design, engineering and manufacturing of energy systems and high-grade, fiber-reinforced composite structures for aerospace, shipbuilding, wind energy, and other applications. MARITCORP employs approximately 1300 employees and has sales of approximately €200 million.

Incentive plan

In 1997, MARITCORP implemented an incentive plan for 160 higher-level subordinates. Prior to this implementation, these subordinates were paid a fixed wage, which was predominately determined by seniority. The firm stated that the incentive plan served the following four related purposes: (1) create clear responsibilities and performance-oriented behavior, (2) promotion based on performance, skills and competencies, (3) employee differentiation, and (4) performance-based compensation. That is, the firm's goal was to give employees incentives to provide (additional) effort and to be able to promote the 'above average' performers.

The firm's incentive plan consists of an annual bonus plan, where the annual bonus is determined by two performance ratings; one based on objective performance measures and one based on subjective performance measures. Although I use the labels 'objective' and 'subjective' performance measures, the firm labeled these measures respectively 'economic' performance measures and 'organizational' performance measures. The human resource manager of the firm corroborated my 'alternative' classification and indicated that it was an accurate reflection of the actual measures used. Examples of the objective performance measures actually used are (1) '2% cost decrease', (2) 'performance compared to budget', and (3) '5% reduction in absenteeism'. Examples of the subjective performance measures actually used are (1) 'improve quality of service', (2) 'good use of resources', (3) 'exposure and availability of products/services', and (4) 'adequate planning'.

The annual bonus plan can best be characterized by the following sharing rule $s(\cdot)$

$$s(\cdot) = \alpha + \beta_1 \left(\sum_{i=1}^{m} b_i \cdot \mathrm{OPM}_i \right) + \beta_2 \left(\sum_{j=1}^{n} c_j \cdot \mathrm{SPM}_j \right)$$

where $\alpha =$ annual salary, $\beta_1 =$ explicit weight put on objective performance, $\beta_2 =$ explicit weight put on subjective performance, $m =$ number of objective performance measures, $n =$ number of subjective performance measures, $b_i =$ implicit weight put on objective performance measure (OPM) i, $c_j =$ implicit weight put on subjective performance measure (SPM) j.

[2] These foreign subsidiaries are not taken into account in this study.

For each employee, the specifics of the annual bonus plan are determined on an annual basis and involve the following steps. First, the size of the total annual bonus as a percentage of annual salary, i.e., $\sum_{w=1}^{2} \beta_w$, is determined by the human resource department and is based on the individual's job, education, and experience. Second, the employee's direct superior allocates the size of the total annual bonus to the two performance dimensions, i.e., β_1 and β_2. Third, the superior chooses the number of measures in each performance dimension, i.e., m and n. Finally, each performance dimension is evaluated ex post, where b_i and c_j are implicitly chosen, and the actual annual bonus is split into two different payments; one based on 'objective' performance and one based on 'subjective' performance. Performance ratings for these dimensions are made on a scale from zero to 100%, where 100% means full bonus payment. It is important to note here that the firm has no explicit rule regarding a forced distribution of the performance ratings. [3] That is, superiors are not obliged to distribute the performance ratings among their subordinates according to a predefined rule. In discussions with the human resource manager of the firm, it became clear that such a rule was also not implicitly communicated within the firm.

Data

The firm provided me with proprietary archival data related to the plan's second year, i.e., 1998. Data are available with respect to the (1) compensation risk (total, objective, and subjective), (2) number of performance measures (total, objective, and subjective), and (3) bonus-related performance ratings (objective and subjective). Data on all variables are available for 124 subordinates. [4] The mean total compensation risk, measured by the *target* annual bonus divided by annual salary, is 8.2%, i.e., approximately one month's salary. Although this percentage might seem low, it is a significant incentive for MARITCORP's subordinates given that the firm moved from no performance evaluation and fixed pay to performance-based compensation. The mean compensation risk based on objective performance measures equals 4.4% of salary, while the mean compensation risk based on subjective performance measures equals 3.8%. The mean percentage point difference between the subjective and objective compensation risk, which reflects the relative use of subjectivity for incentive purposes, equals −0.5% and is significantly different from zero ($p < 0.01$ two-tailed). This means that, on average, superiors rely more on objective performance measures relative to subjective performance measures for incentive purposes. This result is consistent with previous research, which finds that superiors have a general tendency to rely more on objective performance measures (e.g., Ittner et al., 2003).

The mean total number of performance measures used is 4.7, which consists of 2.3 objective performance measures and 2.5 subjective performance measures. The difference between the number of objective and subjective performance measures is statistically significant ($p < 0.05$ two-tailed), which indicates that, on average, superiors use more subjective performance measures than objective performance measures. These results suggest that the number of performance measures used in a performance dimension does *not* reflect the relative importance of that dimension for incentive purposes.

The mean overall performance rating is 61.6%, which means that, on average, 61.6% of the target annual bonus is actually paid. The mean objective (subjective) performance rating equals 62.2% (60.3%). Summary statistics for all variables used in this paper are presented in Table 1 and the correlation between the variables is presented in Table 2.

Measurement issues and empirical specification

In order to test to what extent diversity and subjectivity lead to higher performance ratings (leniency) and more compression of performance

[3] A typical example of a forced distribution is the 15-70-15 rule, which means that superiors are forced to give a 'poor' rating to at least 15% of their subordinates and an 'excellent' rating to another 15%.

[4] The data provided do not relate to all employees included in the incentive plan because some compensation documents received *by* the human resource department were incomplete.

Table 1
Descriptive statistics

Variable	Mean	SD	Actual range	*n*
Total compensation risk	8.2%	2.1%	5.00–15.00	124
Compensation risk based on objective performance measures	4.4%	1.4%	1.40–9.38	124
Compensation risk based on subjective performance measures	3.8%	1.1%	1.50–7.00	124
Total # of performance measures	4.7	1.5	2.00–9.00	124
# of objective performance measures	2.3	0.8	1.00–5.00	124
# of subjective performance measures	2.5	0.9	1.00–6.00	124
Overall performance rating	61.6%	24.2%	0.00–100	124
Objective performance rating	62.2%	29.2%	0.00–100	124
Subjective performance rating	60.3%	26.5%	0.00–100	124

ratings (less differentiation), I use two regression models. The first model relates to the performance rating (leniency) and has the following specification.

$$\text{RATING}_{ij} = \alpha_0 + \alpha_1 D_{j=2} + \sum_{j=1}^{2} \alpha_{2j} D_j \times D_{\text{Diversity-}i,j} + \alpha_3 \sum_{j=1}^{2} D_j \times \text{Comprisk}_{ij} + \varepsilon_{ij} \quad (1)$$

where i relates to subordinates ($i = 1, \ldots, 124$) and j relates to the performance dimension ($j = 1 =$ objective, $j = 2 =$ subjective);

RATING_{ij} performance rating of subordinate i on performance dimension j;
D_j dummy variable that equals 1 if the observation relates to performance dimension j and zero otherwise;
$D_{\text{Diversity-}i,j}$ dummy variable that equals 1 if multiple performance measures are used for subordinate i on performance dimension j and zero otherwise;
Comprisk_{ij} target annual bonus, as a percentage of salary, for subordinate i on performance dimension j.

This regression model examines whether (1) performance ratings on the subjective dimension are, on average, different from the performance ratings on the objective dimension (α_1), (2) performance measure diversity in the objective dimension affects performance ratings (α_{21}), and (3) performance measure diversity in the subjective

Table 2
Correlation between variables

	Total compensation risk	Objective compensation risk	Subjective compensation risk	Total # of performance measures	# of objective performance measures	# of subjective performance measures	Overall performance rating	Objective performance rating
Objective compensation risk	0.845***							
Subjective compensation risk	0.738***	0.263***						
Total # of performance measures	0.234***	0.090	0.309***					
# of objective performance measures	0.153*	0.146	0.092	0.802***				
# of subjective performance measures	0.231**	0.015	0.398***	0.863***	0.389***			
Overall performance rating	0.180**	0.158*	0.125	0.137	0.122	0.108		
Objective performance rating	0.181**	0.202**	0.072	0.186**	0.188**	0.128	0.890***	
Subjective performance rating	0.124	0.113	0.082	0.035	0.028	0.031	0.822***	0.498***

***, **, * is statistically significant at respectively the 1%, 5%, and 10% level (two-tailed).

dimension affects performance ratings (α_{22}), after controlling for the effect of incentives on performance ratings (α_3). Subjectivity provides the superior with discretion and I expect that they use this discretion to give more lenient performance ratings. Evidence consistent with this expectation is found when $\alpha_1 > 0$ because there should, *on average*, be no difference between subordinates' performance ratings on the subjective dimension and on the objective dimension, after controlling for differences in incentives, unless the ratings are biased. Similarly, using multiple objective performance measures provides the superior with discretion, while using multiple subjective performance measures has no such effect because subjectivity per se already provides the superior with discretion. Evidence consistent with leniency, due to diversity, is thus found when $\alpha_{21} > 0$ and $\alpha_{22} = 0$. [5] That is, using multiple objective performance measures leads to higher performance ratings, while this effect is absent when using multiple subjective performance measures. Note that it is important that both $\alpha_{22} > 0$ and $\alpha_{22} = 0$ in order to rule out alternative explanations. If $\alpha_{21} > 0$ but also $\alpha_{22} > 0$, then it is possible that the use of multiple performance measures leads to higher performance ratings because they provide, for example, more effective incentives.

The second model relates to the compression of performance ratings and has the following specification.

$$\begin{aligned}\text{AD\_RATING}_{ij} &= \beta_0 + \beta_1 D_{j=2} + \sum_{j=1}^{2} \beta_{2j} D_j \times D_{\text{Diversity-}i,j} \\ &\quad + \beta_3 \sum_{j=1}^{2} D_j \times \text{AD\_Comprisk}_{ij} + \varepsilon_{ij} \qquad (2)\end{aligned}$$

where i relates to subordinates ($i = 1, \ldots, 124$) and j relates to the performance dimension ($j = 1 =$ objective, $j = 2 =$ subjective);

AD_RATING_{ij} absolute deviation of subordinate i's performance rating on dimension j from the median performance rating on dimension j;

AD_Comprisk_{ij} absolute deviation of subordinate i's compensation risk on dimension j from the median compensation risk on dimension j.

All other variables are as defined above. I measure the compression of performance ratings by calculating the absolute deviation of subordinates' performance ratings on the objective (subjective) dimension from the median performance rating on the objective (subjective) dimension. The smaller the absolute deviation the more subordinates' performance ratings equal the median rating and therefore the more compressed (less differentiated) the ratings are. The second model examines whether (1) the compression of performance ratings on the subjective dimension is, on average, different from the compression of performance ratings on the objective dimension (β_1), (2) performance measure diversity in the objective dimension affects compression of performance ratings (β_{21}), and (3) performance measure diversity in the subjective dimension affects compression of performance ratings (β_{22}), controlling for the effect of incentives (β_3). Following the line of reasoning above, I expect that the performance ratings are, on average, more compressed in the subjective dimension than in the objective dimension (i.e., $\beta_1 < 0$) and the use of multiple performance measures leads to more compression in the objective dimension but not in the subjective dimension (i.e., $\beta_{21} < 0$ and $\beta_{22} = 0$). Once again, it is important that both $\beta_{21} < 0$ and $\beta_{22} = 0$ in order to rule out alternative explanations. If $\beta_{21} < 0$ but also $\beta_{22} < 0$, then it is possible that the use of multiple performance measures leads to more 'compressed' performance ratings because of, for example, error aggregation.

The expectations for both models are summarized in Table 3. In the remainder of this paper, I use the labels $D_{\text{Subjectivity}}$ and $D_{\text{Diversity-obj}}$ ($D_{\text{Diversity-subj}}$) for respectively $D_{j=2}$ and $D_{j=1} \times D_{\text{Diversity-}i,j=1}$ ($D_{j=2} \times D_{\text{Diversity-}i,j=2}$), for ease of exposition.

[5] Strictly speaking, the statistical test for α_{22} relates to rejecting the null-hypothesis of $\alpha_{22} > 0$ in favor of the alternative hypothesis of $\alpha_{22} \leqslant 0$, since it is statistically impossible to accept the "traditional" null-hypothesis of $\alpha_{22} = 0$. This line of reasoning applies to all coefficients mentioned in the paper that are expected to be zero.

Table 3
Summary of predictions

	Leniency	Compression
Subjectivity	$\alpha_1 > 0$	$\beta_1 < 0$
Diversity	$\alpha_{21} > 0$ and $\alpha_{22} = 0$	$\beta_{21} < 0$ and $\beta_{22} = 0$

Regression specifications:

$$\text{RATING}_{ij} = \alpha_0 + \alpha_1 D_{j=2} + \sum_{j=1}^{2} \alpha_{2j} D_j \times D_{\text{Diversity-}i,j} + \alpha_3 \sum_{j=1}^{2} D_j \times \text{Comprisk}_{ij} + \varepsilon_{ij}$$

$$\text{AD\_RATING}_{ij} = \beta_0 + \beta_1 D_{j=2} + \sum_{j=1}^{2} \beta_{2j} D_j \times D_{\text{Diversity-}i,j} + \beta_3 \sum_{j=1}^{2} D_j \times \text{AD\_Comprisk}_{ij} + \varepsilon_{ij}$$

RATING_{ij}: performance rating of subordinate i on performance dimension j.
AD_RATING_{ij}: absolute deviation of subordinate i's performance rating on dimension j from the median performance rating on dimension j.
D_j: dummy variable that equals 1 if the observation relates to performance dimension j and zero otherwise.
$D_{\text{Diversity-}i,j}$: dummy variable that equals 1 if multiple performance measures are used for subordinate i on performance dimension j and zero otherwise.
Comprisk_{ij}: target annual bonus, as a percentage of salary, for subordinate i on performance dimension j.
AD_Comprisk_{ij}: absolute deviation of subordinate i's compensation risk on dimension j from the median compensation risk on dimension j.

I use Tobit regression, instead of standard OLS regression, to estimate model (1) and (2) because the dependent variables are censored. The Tobit model can be defined as a latent variable model and is generally characterized by (Greene, 2000):

$$y_i^* = \boldsymbol{\beta}' \mathbf{x}_i + \varepsilon_i$$
$$y_i = a \quad \text{if } y_i^* \leqslant a;$$
$$y_i = b \quad \text{if } y_i^* \geqslant b; \text{ and}$$
$$y_i = y_i^* \quad \text{otherwise}$$

where a and b are constants and the latent variable y_i^* satisfies the classical linear model assumptions.

Results

The results of the Tobit regressions for model (1) and (2) are presented in Table 4. [6] The results for model (1) show that $D_{\text{Subjectivity}}$ has a significant positive effect on RATING ($p < 0.10$ two-tailed). That is, the performance ratings on the subjective dimension are, on average, higher than the performance ratings on the objective dimension, which suggests that superiors give more lenient performance ratings when using subjectivity. Further, $D_{\text{Diversity-obj}}$ has a significant positive effect on RATING ($p < 0.01$ two-tailed), which indicates that the use of multiple objective performance measures leads to higher performance ratings. $D_{\text{Diversity-subj}}$, on the other hand, is not related to RATING and the use of multiple subjective performance measures has thus no effect on performance ratings. The combined results for $D_{\text{Diversity-obj}}$ and $D_{\text{Diversity-subj}}$ suggest that superiors give more lenient performance ratings when using multiple objective performance measures, which is consistent with expectations. The control variable Comprisk has a significant positive effect on RATING ($p < 0.10$ two-tailed), which suggests that incentives have a positive impact on performance. Finally, model (1) as a whole is significant, as reflected by the significant Log likelihood ratio ($p < 0.05$ two-tailed).

Overall, the results for model (1) are consistent with the expectation that superiors give more lenient performance ratings when they have discretion in performance evaluation, where discretion is either due to the use of subjectivity per se or the

[6] Specification tests indicate that the two Tobit models do not exhibit problems of multicollinearity and heteroskedasticity. However, the null-hypothesis of normally distributed error terms is rejected for both models. To examine whether the results are sensitive to this deviation from normality, I perform a bootstrapping procedure by resampling the observations. The results indicate that the findings presented in Table 4 are robust.

Table 4
Tobit regressions of the impact of performance measure diversity and subjectivity on the performance rating and the absolute deviation of the performance rating from the median rating (z-statistics are in parentheses)

Independent variable		RATING		AD_RATING
Constant	α_0	0.33***	β_0	0.37***
		(3.37)		(10.79)
$D_{\text{Subjectivity}}$	α_1	0.17*	β_1	−0.09*
		(1.84)		(1.79)
$D_{\text{Diversity-obj}}$	α_{21}	0.19***	β_{21}	−0.16***
		(2.94)		(4.31)
$D_{\text{Diversity-subj}}$	α_{22}	−0.02	β_{22}	−0.06
		(0.30)		(1.30)
Comprisk	α_3	0.03*		
		(1.83)		
AD_Comprisk			β_3	−0.00
				(0.21)
# of censored observation		46		18
Total # of observations		248		248
Log likelihood ratio		11.10**		13.77***

***, **, * is statistically significant at respectively the 1%, 5%, and 10% level (two-tailed).
Regression specifications:

$$\text{RATING}_{ij} = \alpha_0 + \alpha_1 D_{j=2} + \sum_{j=1}^{2} \alpha_{2j} D_j \times D_{\text{Diversity-}i,j} + \alpha_3 \sum_{j=1}^{2} D_j \times \text{Comprisk}_{ij} + \varepsilon_{ij}$$

$$\text{AD\_RATING}_{ij} = \beta_0 + \beta_1 D_{j=2} + \sum_{j=1}^{2} \beta_{2j} D_j \times D_{\text{Diversity-}i,j} + \beta_3 \sum_{j=1}^{2} D_j \times \text{AD\_Comprisk}_{ij} + \varepsilon_{ij}$$

RATING_{ij}: performance rating of subordinate i on performance dimension j.
AD_RATING_{ij}: absolute deviation of subordinate i's performance rating on dimension j from the median performance rating on dimension j.
D_j: dummy variable that equals 1 if the observation relates to performance dimension j and zero otherwise.
$D_{\text{Diversity-}i,j}$: dummy variable that equals 1 if multiple performance measures are used for subordinate i on performance dimension j and zero otherwise.
Comprisk_{ij}: target annual bonus, as a percentage of salary, for subordinate i on performance dimension j.
AD_Comprisk_{ij}: absolute deviation of subordinate i's compensation risk on dimension j from the median compensation risk on dimension j.

use of multiple objective performance measures. The data therefore provide support for hypotheses 1A and 1B.

Although the results are *statistically* significant and in the expected direction, the next question is whether they are *economically* significant. That is, is the impact of subjectivity and diversity on the performance rating, and thus on the actual bonus payout, large enough to have economic consequences? [7] In order to test the economic significance, I estimate the marginal effects of $D_{\text{Subjectivity}}$ and $D_{\text{Diversity-obj}}$ on the performance rating. In Tobit regression, the regression coefficients represent the marginal effects of the independent variables on the (unobserved) latent variable. The variable of interest, however, is the *observed* dependent variable. The marginal effect of an independent variable on the observed dependent variable can be estimated by (Greene, 2000): $\beta \times \text{Prob}[a < y_i^* < b]$, where β is the Tobit regression coefficient.

The marginal effect of $D_{\text{Subjectivity}}$ equals 0.14, which means that the percentage point difference between the subjective performance rating and the objective performance rating is, on average, 14%. This suggests that an additional 14% of the target annual bonus based on subjective performance is

[7] The term 'economic consequences' should be interpreted broadly and includes, for example, the potential effects on (future) subordinates' behavior.

paid compared to the bonus payout-percentage based on objective performance. The marginal effect of $D_{\text{Diversity-obj}}$ equals 0.16. This suggests that subordinates evaluated on multiple objective performance measures receive an additional 16% of their target annual bonus based on objective performance compared to subordinates evaluated on a single objective performance measure. Although there is no formal rule to determine economic significance, the marginal effects of both $D_{\text{Subjectivity}}$ and $D_{\text{Diversity-obj}}$ on the performance ratings seem to be large enough to have economic consequences.

Model (2), which examines the compression of performance ratings, shows results similar to those for model (1). First, $D_{\text{Subjectivity}}$ has a significant negative effect on AD_RATING ($p < 0.10$ two-tailed), which indicates that, on average, the performance ratings on the subjective dimension are closer to the median rating than the performance ratings on the objective dimension. This result is consistent with the expectation that superiors compress the performance ratings more when using subjectivity. Second, regarding the impact of performance measure diversity, the results indicate that $D_{\text{Diversity-obj}}$ has a significant negative effect on AD_RATING ($p < 0.01$ two-tailed), while $D_{\text{Diversity-subj}}$ is not related to AD_RATING. These findings suggest that superiors give more compressed performance ratings when using multiple objective performance measures. Finally, the control variable AD_Comprisk is not related to AD_RATING, while the model as a whole is significant ($p < 0.01$ two-tailed).

Overall, the results for model (2) are consistent with the expectation that superiors give more compressed performance ratings when they have discretion in performance evaluation, where discretion is either due to the use of subjectivity per se or the use of multiple objective performance measures. The data therefore provide support for hypotheses 2A and 2B.

Additional tests

As noted above, the results for model (1) and (2) are consistent with expectations. However, both models (1) and (2) use two observations per subordinate and, as a result, the observations are not independent. To control for this lack of independence, I perform an additional test based on differences. More specifically, I run the following two OLS regressions.

$$\begin{aligned}\Delta\text{RATING}_i = {} & \delta_0 + \delta_1 D_{\text{Diversity-obj-}i} \\ & + \delta_2 D_{\text{Diversity-subj-}i} \\ & + \delta_3 \Delta\text{Comprisk}_i + \varepsilon_i \end{aligned} \quad (3)$$

$$\begin{aligned}\Delta\text{AD\_RATING}_i = {} & \gamma_0 + \gamma_1 D_{\text{Diversity-obj-}i} \\ & + \gamma_2 D_{\text{Diversity-subj-}i} \\ & + \gamma_3 \Delta\text{AD\_Comprisk}_i + \varepsilon_i \end{aligned} \quad (4)$$

where

ΔRATING_i difference between subordinate i's performance rating on the objective dimension and the subjective dimension;

$\Delta\text{AD\_RATING}_i$ difference between the absolute deviation of subordinate i's performance rating from the median performance rating on the objective dimension and the absolute deviation of his/her performance rating from the median performance rating on the subjective dimension;

$D_{\text{Diversity-obj-}i}$ dummy variable that equals 1 if multiple objective performance measures are used for subordinate i and zero otherwise;

$D_{\text{Diversity-subj-}i}$ dummy variable that equals 1 if multiple subjective performance measures are used for subordinate i and zero otherwise;

$\Delta\text{Comprisk}_i$ difference between subordinate i's target annual bonus, as a percentage of salary, based on the objective performance dimension and the subjective performance dimension;

$\Delta\text{AD\_Comprisk}_i$ difference between the absolute deviation of subordinate i's compensation risk from the median compensation risk on the objective dimension and the absolute deviation of his/her compensation risk from the median compensation risk on the subjective dimension.

Evidence consistent with superiors giving more lenient (compressed) performance ratings when using subjectivity is found when $\delta_0 < 0$ ($\gamma_0 > 0$).

Table 5
OLS regression of the impact of performance measure diversity and subjectivity on differences in the performance rating and the absolute deviation of the performance rating from the median rating between the objective and subjective performance dimension (*t*-statistics are in parentheses)

Independent variable	ΔRATING	ΔAD_RATING
Constant	−0.20**	0.19***
	(2.40)	(3.44)
$D_{\text{Diversity-obj}}$	0.12*	−0.09*
	(1.71)	(1.97)
$D_{\text{Diversity-subj}}$	0.13	−0.11**
	(1.55)	(2.13)
ΔComprisk	0.02	
	(1.06)	
ΔAD_Comprisk		−0.01
		(0.60)
# of observations	124	124
Adjusted *R*-square	0.04	0.06
F-statistic	2.86**	3.63**

***, **, * is statistically significant at respectively the 1%, 5%, and 10% level (two-tailed).
ΔRATING: difference between a subordinate's performance rating on the objective dimension and the subjective dimension.
ΔAD_RATING: difference between the absolute deviation of a subordinate's performance rating from the median performance rating on the objective dimension and the absolute deviation of his/her performance rating from the median performance rating on the subjective dimension.
$D_{\text{Diversity-obj}}$: dummy variable that equals 1 if multiple objective performance measures are used and zero otherwise.
$D_{\text{Diversity-subj}}$: dummy variable that equals 1 if multiple subjective performance measures are used and zero otherwise.
ΔComprisk: difference between the target annual bonus, as a percentage of salary, based on the objective performance dimension and the subjective performance dimension.
ΔAD_Comprisk: difference between the absolute deviation of a subordinate's compensation risk from the median compensation risk on the objective dimension and the absolute deviation of his/her compensation risk from the median compensation risk on the subjective dimension.

Similarly, evidence consistent with superiors giving more lenient (compressed) performance ratings when using multiple objective performance measures is found when $\delta_1 > 0$ and $\delta_2 = 0$ ($\gamma_1 < 0$ and $\gamma_2 = 0$).

The results of the OLS regressions for models (3) and (4) are presented in Table 5 and are consistent with those provided in Table 4. [8] The subjective performance dimension is, on average, characterized by higher performance ratings compared to the objective performance dimension ($p < 0.05$ two-tailed) and lower deviations from the median performance rating ($p < 0.01$ two-tailed). Further, the use of multiple objective performance measures increases the difference between the performance ratings on the objective dimension and the subjective dimension ($p < 0.10$ two-tailed) and decreases the difference between the deviations from the median rating on both dimensions ($p < 0.10$ two-tailed). Finally, the use of multiple subjective performance measures is not related to the difference between the performance ratings on the two dimensions and decreases the difference between the deviations from the median rating on both dimensions ($p < 0.05$ two-tailed). Although this latter result is significant, the sign of the coefficient for $D_{\text{Diversity-subj}}$ is not opposite to that for $D_{\text{Diversity-obj}}$, which is consistent with a bias story.

In sum, the results presented in this section provide strong support for the prediction that when superiors have discretion in performance evaluation, either due to the use of subjectivity per se or the use of multiple objective performance

[8] Specification tests indicate that the two OLS models do not exhibit problems of multicollinearity but do exhibit heteroskedasticity and marginal deviations from normality. To test the robustness of the results, I perform a bootstrapping procedure by resampling the observations, which is valid in the presence of heteroskedasticity (Davidson & MacKinnon, 1993, pp. 767–768). The bootstrapped results are consistent with those presented in Table 5.

measures, they significantly bias the evaluation in the sense that they give more lenient and more compressed performance ratings.

Discussion and conclusion

The empirical results presented in the previous section have some important implications for scorecard-type of performance measurement and reward systems. These systems are characterized by multiple performance measures (diversity) and an increased use of subjectivity (e.g., Ittner et al., 2003). Furthermore, the purpose of these systems is to motivate employees to improve performance and to differentiate among employees based on their ability and skills so as to make better promotion decisions. Previous empirical evidence indicates that when multiple performance measures are available and superiors have discretion in 'weighting' the different performance measures, there is a general tendency to put more weight on objective and common measures of performance, which results in less 'balance' (e.g., Ittner et al., 2003; Lipe & Salterio, 2000). The empirical results in this paper indicate an additional problem with the use of multiple performance measures and subjectivity. First, the results show that performance measure diversity leads to more lenient performance ratings and less differentiation among employees. Kaplan and Norton (2001) claim that, although the Balanced Scorecard should be used as a 'strategic management system', it can be linked to incentive compensation. They state that 'compensation can be based on 25 strategic measures' without causing problems (2001, p. 152). However, this is more than five times the mean number of performance measures and almost three times the maximum number of performance measures used in the research site examined in this paper. It is therefore questionable whether a Balanced Scorecard that includes a large number of performance measures is effective as a performance measurement and reward system, if the superior has discretion in weighting these measures. [9]

Second, Kaplan and Norton (1996) state that a balanced scorecard with multiple performance measures makes the use of subjectivity for incentive purposes easier. However, the empirical results indicate that subjectivity leads to bias in performance evaluation. That is, if more subjectivity is used in evaluating and rewarding employees, superiors give higher performance ratings and compress these ratings. As a result, the firm is unable to separate the highly skilled employees from the less skilled employees. If skills and competencies are important determinants of promotions, then the use of subjective performance measures makes these promotion decisions more difficult.

As with any empirical study, this study has its limitations. First, the data do not allow me to examine the behavior of individual superiors and the analysis therefore assumes that, on average, all superiors behave in an identical way. Although research in psychology indicates that superiors have a general tendency to bias the performance ratings, in terms of leniency and compression, it might be that superior-specific characteristics influence bias and future research can address these issues. Second, because the data are cross-sectional data of a single year, I am unable to examine to what extent performance evaluation bias persists and whether the firm actually incurs the assumed indirect costs of bias. Although there is some evidence that superiors are persistent in their evaluation behavior (e.g., Kane, Bernardin, Villanova, & Peyrefitte, 1995), an opportunity for future research is to gather time-series data to examine whether this finding is generalizable and to test to what extent performance evaluation bias affects, for example, motivation. Finally, the data relate to a single firm. Using data from a single firm controls for 'other' factors that can affect performance evaluation bias but it reduces the generalizability of the results. Future research can examine to what extent the findings in this paper are generalizable by gathering data from multiple firms.

[9] It should be noted that a performance measurement system that includes a large number of performance measures could be effective for purposes other than providing incentives. For example, Ittner, Larcker, and Randall (2003) examine a broad set of measurement system uses, which include goal setting, capital investment decisions, problem identification, performance evaluation, and external disclosure. They find that overall measurement diversity is positively related to measurement system satisfaction and stock returns.

Acknowledgements

I gratefully appreciate the support of the firm that provided me with the data for this study, especially the human resource manager of the firm. I am further grateful to Ken Merchant, Erik Peek, Mike Shields, Dhinu Srinivasan, an anonymous reviewer, and seminar participants at Tilburg University, Maastricht University, the 2002 AAA Annual Meeting in San Antonio, and the Production, Use, and Maintenance of Human Capital Conference in Maastricht for their comments and suggestions.

References

Baiman, S., & Rajan, M. V. (1995). The informational advantage of discretionary bonus schemes. *The Accounting Review, 70*, 557–579.

Baker, G., Gibbons, R., & Murphy, K. J. (1994). Subjective performance measures in optimal incentive contracts. *Quarterly Journal of Economics, 109*, 1125–1156.

Datar, S., Cohen Kulp, S., & Lambert, R. A. (2001). Balancing performance measures. *Journal of Accounting Research, 39*, 75–92.

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics*. New York, NY: Oxford University Press.

Feltham, G. A., & Xie, J. (1994). Performance measure congruity and diversity in multi-task principal/agent relations. *The Accounting Review, 69*, 429–453.

Gibbs, M., Merchant, K. A., Van der Stede, W. A., & Vargus, M. E. (2002). *Causes and effects of subjectivity in incentives*. Working Paper. University of Chicago/University of Southern California/University of Texas at Dallas.

Govindarajan, V. (1984). Appropriateness of accounting data in performance evaluation: an empirical examination of environmental uncertainty as an intervening variable. *Accounting, Organizations and Society, 9*, 125–135.

Govindarajan, V., & Gupta, A. K. (1985). Linking control systems to business unit strategy: impact on performance. *Accounting, Organizations and Society, 10*, 51–66.

Greene, W. H. (2000). *Econometric analysis*. Upper Saddle River, NJ: Prentice Hall.

Hayes, R. M., & Schaefer, S. (2000). Implicit contracts and the explanatory power of top executive compensation for future performance. *Rand Journal of Economics, 31*, 273–293.

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

Ittner, C. D., & Larcker, D. F. (1998). Innovations in performance measurement: trends and research implications. *Journal of Management Accounting Research, 10*, 205–238.

Ittner, C. D., & Larcker, D. F. (1999). *The effects of performance measure diversity on incentive plan outcomes*. Working Paper. University of Pennsylvania.

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

Ittner, C. D., Larcker, D. F., & Rajan, M. V. (1997). The choice of performance measures in annual bonus contacts. *The Accounting Review, 72*, 231–255.

Ittner, C. D., Larcker, D. F., & Randall, T. (2003). Performance implications of strategic performance measurement in financial services firms. *Accounting, Organizations and Society, 28*, 715–741.

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: the performance appraisal purpose effect. *Personnel Psychology, 50*, 905–925.

Kane, J. S., Bernardin, H. J., Villanova, P., & Peyrefitte, J. (1995). Stability of rater leniency: three studies. *Academy of Management Journal, 38*, 1036–1051.

Kaplan, R. S., & Norton, D. P. (1996). *The balanced scorecard: translating strategy into action*. Boston, MA: Harvard Business School Press.

Kaplan, R. S., & Norton, D. P. (2001). Transforming the balanced scorecard from performance measurement to strategic management: part II. *Accounting Horizons, 15*, 147–160.

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

Milkovich, G. T., & Newman, J. M. (1993). *Compensation*. Homewood, IL: Irwin.

Murphy, K. J., & Oyer, P. (2001). *Discretion in executive incentive contracts: theory and evidence*. Working Paper. University of Southern California/Stanford University.

Prendergast, C., & Topel, R. (1993). Discretion and bias in performance evaluation. *European Economic Review, 37*, 355–365.

Prendergast, C., & Topel, R. (1996). Favoritism in organizations. *Journal of Political Economy, 104*, 958–978.

Wallace, J. S. (1997). Adopting residual income-based compensation plans: do you get what you pay for? *Journal of Accounting and Economics, 24*, 275–300.