European Accounting Review

ISSN: 0963-8180 (Print) 1468-4497 (Online) Journal homepage: www.tandfonline.com/journals/rear20

Differential Weighting of Objective Versus Subjective Measures in Performance Evaluation: Experimental Evidence

Narisa Tianjing Dai, Xi (Jason) Kuang & Guliang Tang

To cite this article: Narisa Tianjing Dai, Xi (Jason) Kuang & Guliang Tang (2018) Differential Weighting of Objective Versus Subjective Measures in Performance Evaluation: Experimental Evidence, European Accounting Review, 27:1, 129-148, DOI: 10.1080/09638180.2016.1234402

To link to this article: https://doi.org/10.1080/09638180.2016.1234402

European Accounting Review, 2018
Vol. 27, No. 1, 129–148, https://doi.org/10.1080/09638180.2016.1234402

Differential Weighting of Objective Versus Subjective Measures in Performance Evaluation: Experimental Evidence

NARISA TIANJING DAI*, XI (JASON) KUANG** and GULIANG TANG*

**Business School, University of International Business and Economics, Beijing, P.R. China, **Scheller College of Business, Georgia Institute of Technology, Atlanta, USA*

(*Received: June 2015; accepted: August 2016*)

Abstract In this paper, we conduct two experiments to investigate how managers' differential weighting of objective versus subjective measures affects their performance-evaluation decisions. Drawing on psychological theory, we predict that managers heuristically perceive objective measures to be more scientific than subjective measures. As a result, their performance-evaluation decisions are influenced more by objective measures than by subjective measures. Experimental results are consistent with our prediction. Supplemental analyses further support our theory by showing that participants do not perceive objective measures to be more important for performance evaluation nor do they perceive subjective measurement to be inappropriate. The implications of our findings for management accounting research and practice are discussed.

1. Introduction

Many firms use both objective and subjective performance measures in evaluation and compensation practices (Ittner & Larcker, 1998; Ittner, Larcker, & Meyer, 2003; Luft, 2009; Van der Stede, Chow, & Lin, 2006).[1] Subjective evaluation is valuable because it allows firms to utilize relevant but non-contractible information to more accurately assess employees' performance (Baiman & Rajan, 1995; Baker, Gibbons, & Murphy, 1994; Feltham & Xie, 1994; Rajan & Reichelstein, 2009). However, such benefits may not materialize because the effectiveness of subjective evaluation can be impaired by evaluators' opportunism (e.g. favoritism) or cognitive limitations (Bol, 2008). The purpose of this paper is to investigate the behavioral effects of managers' weighting of objective versus subjective measures on their final performance-appraisal decisions.

Correspondence Address: Xi (Jason) Kuang, Scheller College of Business, Georgia Institute of Technology, 800 West Peachtree Street NW, Atlanta, GA30308-1149, USA. Email: jason.kuang@scheller.gatech.edu

Paper accepted by Victor Maas.

[1]Consistent with the accounting literature (e.g., Gibbs et al., 2004; Woods, 2012), in this paper "objective measures" refer to performance measures that are directly quantifiable using existing data records, and "subjective measures" refer to performance measures that are not quantifiable from data records and can only be determined based on human perceptions or judgments.

Research suggests that managers often make decisions based on heuristics rather than rational analysis (Beach & Mitchell, 1978; Sadler-Smith, 2004). Specifically, managers tend to apply previously acquired stereotypic knowledge or beliefs to solving a current decision problem, with the goal of saving cognitive effort (Dane & Pratt, 2007; Schwenk, 1988). However, in doing so, managers may fail to incorporate relevant information from the current environment, which can affect the quality of the final decision (Ireland, Hitt, Bettis, & Auld de Porras, 1987; Walsh, 1995). For our purpose, the key difference between objective and subjective performance measures is whether the measure is based on data supplied by information systems or based on human judgments. Managers may assign more weight to objective measures than subjective measures because people generally believe that 'hard' non-human attributes are more scientific than 'soft' attributes involving human judgments (Hsee, Zhang, Yu, & Xi, 2003; Tsai & Hsee, 2009). As a result, managers' performance-evaluation decisions are more likely to be influenced by objective measures than by subjective measures.

We conduct two laboratory experiments to test our prediction. In both experiments, professional employees act as the Human Resources Director of a corporation and rate the performance of two division managers. Experiment One has a 1×2 design, in which we manipulate within-participant whether one manager performed better on subjective or objective measures than the other manager. That is, one manager performed better on subjective measures, but worse on objective measures, as compared to the other manager (participants are provided with historical information suggesting that objective and subjective measures have the same level of accuracy and reliability). Participants give a performance rating for each of the two managers and we randomly vary the order in which the two managers' performance information is presented to participants. Consistent with psychological theory, we find that the performance rating is significantly higher for the manager who performed better on *objective* measures but worse on *subjective* measures. Also consistent with our theory, post-experimental questionnaire data show that participants' tendency to assign more weight to objective measures than subjective measures is related to the perceived scientific nature of their decisions.

To examine the robustness of the results of Experiment One and give our theory a stricter test, we design a second experiment using the same basic setting. In Experiment Two, the two managers have differential performance on two measures: an innovation measure and a capacity utilization measure. One manager always performed better on the innovation measure but worse on the capacity utilization measure, as compared to the other manager. Experiment Two uses a 1×2 design, in which we manipulate between participants whether the innovation measure is subjective or objective (the capacity utilization measure is always objective). Participants are again told that subjective and objective measures have the same level of accuracy and reliability. As in Experiment One, participants give a performance rating for each of the two managers. We predict and find that the performance rating is significantly *lower* for the manager who performed better on the innovation measure when this measure is subjective (i.e. the manager's superior subjective-measure performance appears to be outweighed by his inferior objective-measure performance), but is not significantly different between the two managers when the innovation measure is *objective* (i.e. the manager's superior performance on one objective measure appears to make up for his inferior performance on another objective measure). Post-experimental questionnaire data show that this observed pattern of performance ratings is mediated by participants' perceptions of the scientific nature of their decisions.

Overall, the results of the two experiments suggest that, consistent with psychological theory, participants rely more on objective measures than subjective measures with the attempt to make scientific performance-evaluation decisions. As a consequence, their performance rating tilts toward objective measures even when there is historical information indicating that objective and subjective measures are equally accurate and reliable. Supplemental analyses show that these

results are not attributable to participants attaching differential importance to objective versus subjective measures or participants questioning the appropriateness of subjective measurement, thus adding credence to our theory.

This study makes several contributions to management accounting research and practice. First, prior accounting research has found that subjective evaluation is likely affected by the evaluator's cognitive limitations (e.g. Bailey, Hecht, & Towry, 2011; Bol & Smith, 2011; Lipe & Salterio, 2000). We extend this line of research by showing that, when evaluating performance based on a mix of objective and subjective measures, evaluators' decisions are disproportionately influenced by objective measures because they heuristically perceive objective measures to be more scientific than subjective measures. Second, we provide an important caveat for research advocating the contracting benefits of subjective measures (Baker et al., 1994; Feltham & Xie, 1994; Rajan & Reichelstein, 2009). Our results suggest that the theoretically derived benefits of subjective measures may not be fully realized if subjective measures are undervalued in the evaluation process. Factors like this should be considered in revising existing models for maximizing the overall effectiveness of performance measurement. Third, our findings have important implications for management control practices. Firms often combine objective and subjective measures in performance evaluation (Jarvis, Curran, Kitching, & Lightfoot, 2000; Kaplan & Norton, 1996; Rich, Bommer, MacKenzie, Podsakoff, & Johnson, 1999). To the extent that subjective measures are valued less than objective measures, employees may opportunistically shift their attention and effort to tasks that are objectively measured. Such shifts, however, can be suboptimal for the firm (Ahn, Hwang, & Kim, 2010). Therefore, firms should understand these potential adverse effects on organizational efficiency and adjust incentive policies accordingly.

The remainder of the paper is organized as follows. The next section lays out the theoretical background and the hypothesis for our study. Section 3 presents the method and results of Experiment One, and Section 4 does the same for Experiment Two. Section 5 reports the results of supplemental analyses. Section 6 discusses our findings and concludes.

2. Background

2.1. Subjective Performance Evaluation

When objective measures only provide noisy signals about employees' real performance, the use of such measures may cause distorted incentives, undue risks for employees, and, in turn, dysfunctional employee behavior (Baker, 2000; Feltham & Xie, 1994; Holmstrom & Milgrom, 1991; Prendergast, 1999). Prior research suggests that subjective evaluation can help mitigate this problem because it enables firms to utilize pertinent information that cannot be objectively measured or contracted upon in performance evaluation, thereby reducing incentive distortion and better aligning the interests of firm and employees (Baiman & Rajan, 1995; Baker, 2000; Budde, 2007; Rajan & Reichelstein, 2009). Consistent with this research, there has been empirical evidence attesting to the contracting benefits of subjective evaluation. For example, Bushman, Indjejikian, and Smith (1996) and Hayes and Schaefer (2000) find more subjectivity in CEO compensation when financial measures are inadequate indicators of firm performance. Subjective evaluation is used more often in private firms (where objective measures are less available than in public firms) (Murphy & Oyer, 2003) and in multinational corporations with higher levels of headquarters-subsidiary interdependence (where output is less controllable by subsidiaries) (Du, Deloof, & Jorissen, 2013). Gibbs, Merchant, Van der Stede, and Vargus (2004) find that subjective evaluation mitigates goal incongruence between firm and employees and helps increase employee satisfaction and performance.

The benefits of subjective evaluation, however, may not materialize because the accuracy of evaluation can be impaired by evaluators' opportunism, such as effort aversion (Baker, Jensen, & Murphy, 1988; Murphy, 1992) and favoritism toward subordinates (Prendergast & Topel, 1996). Along these lines, Moers (2005) shows that the use of subjective measures leads to more compressed and more lenient evaluation results, and Bol (2011) finds that such tendencies are more pronounced when managers face higher information-gathering costs or stronger relationships with employees. Relatedly, Bol, Kramer, and Maas (2016) find that performance-evaluation compression is mitigated as the accuracy of performance information increases, but only when evaluation results are publicly disclosed.

The effectiveness of subjective evaluation also may be affected by evaluators' cognitive limitations, as suggested by prior experimental research. Lipe and Salterio (2000) find that superiors who evaluate multiple subordinates over-rely on common measures, ignoring each subordinate's unique measures (see also Banker, Chang, & Pizzini, 2004; Libby, Salterio, & Webb, 2004). Evaluators' ability to incorporate relevant non-contractible information in performance evaluation is affected by their mental models (Krishnan, Luft, & Shields, 2005), decision approaches (Bailey et al., 2011), and whether the firm uses a formulaic measurement system (Long, Mertins, & Vansant, 2015). In a multi-task setting, Bol and Smith (2011) find that evaluators' subjective rating for a later task is anchored on the objective measure of an unrelated prior task and that evaluators adjust their rating to redress inequity caused by the objective measure.[2] Subjective performance rating is also found to be biased by the superior's perception of the effectiveness of corporate strategy (Johnson, Reckers, & Barlett, 2014) and prior experience with the subordinate (Kramer & Maas, 2016). Other research finds that holding superiors accountable for evaluation results actually increases their suboptimal use of diagnostic versus non-diagnostic performance information (Barlett, Johnson, & Reckers, 2014). Our study extends this experimental literature by examining how differential weighting of objective and subjective measures affects performance-evaluation decisions.

Relevant to our study, Ittner et al. (2003, p. 732) provide archival evidence that objective measures have a greater impact on the overall performance rating than do subjective measures, in settings where subjective measures are 'less accurate and reliable' than objective measures because of the rater's opportunistic behavior such as favoritism and excessive discretion. We extends this research by investigating managers' weighting of subjective versus objective measures when there is evidence that these two types of measures are equally accurate and reliable. We argue that, despite such evidence, managers may still give less weight to subjective measures due to their cognitive biases, as elaborated below.

2.2. Psychological Theory

In our framework, agency theory contends that, if multiple measures are used to evaluate performance, the weights assigned to these measures should depend on the measure's sensitivity (i.e. covariation with the agent's action) and precision (i.e. noisiness) (Banker & Datar, 1989). In our study, participants are provided with historical information suggesting that objective and subjective measures are equally accurate (i.e. sensitive) and reliable (i.e. precise) in reflecting actual performance. Therefore, from the agency theory perspective, objective and subjective measures should be weighted equally in performance evaluation.

However, when making complex decisions, managers may veer from rational analysis and rely instead on heuristics (Boudon, 1998; Hogarth & Kunreuther, 1995; Simon, 1987; Slovic, Fischhoff, & Lichtenstein, 1977) because heuristic reasoning simplifies the decision process and

[2]Woods (2012) reports similar results using field data.

saves cognitive resources (Hoyer, 1984; Ouellette & Wood, 1998). To the extent that managers over-rely on heuristics without adequately incorporating useful information from the dynamic environment (Dane & Pratt, 2007; Kiesler & Sproull, 1982; Sadler-Smith, 2004), their final decision may be inaccurate or suboptimal for the current context (Ireland et al., 1987; Walsh, 1995). For our purpose, despite the presence of information that objective and subjective measures are equally accurate and reliable, managers may make the performance-evaluation decision based on heuristic reasoning, assigning more weight to objective measures. Specifically, people tend to believe that information collected from non-human sources is more 'scientific' than information based on human judgments, regardless of the usefulness of the information. Hsee et al. (2003) refer to this belief as 'lay scientism'. For example, when making a purchase choice between two stereo systems (one has higher power and the other has higher sound richness), most people chose the high-power model when power is measured using a technical parameter and sound richness is measured based on personal experience, even though they indicated that they would enjoy the high-richness model more (Hsee & Hastie, 2006; Hsee et al., 2003).

In our setting, lay scientism may arise because people generally believe that scientific knowledge requires verifiability (Brown, 1988; Dalrymple, 2003; Healy, 1993). Admittedly, compared to objective measures, subjective measures are less self-evident and more difficult to validate using observational evidence. To the extent that people use verifiability as a proxy for the scientific nature of performance measures, over time the notion that subjective measures are non-scientific is created in people's minds and accepted as a default perception (Ford, Kraiger, & Schechtman, 1986; Lawler, 1971; Muckler & Seven, 1992). As discussed earlier, managers often engage in heuristics in decision-making, with the goal of saving cognitive effort. Therefore, when making performance-evaluation decisions, managers may rely on the perceived scientific nature of measures rather than a rational analysis of current decision-relevant information (e.g. evidence about the accuracy and reliability of measures). This heuristic is profoundly ungrounded because empirical studies find that, in many areas, subjective measures are as valid as objective measures in performance evaluation (Hoffman, Nathan, & Holden, 1991; Nathan & Alexander, 1988; Wall et al., 2004), suggesting that 'the "subjective" versus "objective" distinction may be more illusory than real' (Nathan & Alexander, 1988, p. 531).

Based on the proceeding discussion, we predict that evaluators will rely more on objective measures than subjective measures, even if there is evidence indicating that these two types of measures are equally accurate and reliable. As a result, their performance-evaluation decision will tilt toward objective measures. We formally state this in the following hypothesis:

Hypothesis Managers' performance-evaluation decision will be influenced more by objective measures than by subjective measures, *ceteris paribus*.

We conduct two experiments to test our hypothesis. Next, we describe the setting and design of Experiment One, and report the results of hypothesis tests. Then, we do the same for Experiment Two.

3. Experiment One

3.1. Experimental Setting and Design

Participants assume the role of a regional Human Resources Director of XYZ Corporation, a large wholesaler of pharmaceuticals and medical instruments. The strategic objective of XYZ Corporation is to maximize operating profit. The corporation has two distribution centers (referred to as Distribution Center A and Distribution Center B), which are responsible for the

storage and packaging of products and the fulfillment and shipment of customer orders. These two distribution centers are similar in terms of business model, the scale of operation, functionality, and customer base. The current strategic objective of both distribution centers is to provide quality distribution services with high efficiency.

Participants' task is to evaluate the performance of the distribution center managers based on five measures (i.e. on-time delivery, warehouse capability utilization, equipment maintenance and adjustment, emergency and complexity management, and employee ability enhancement). Both managers have the same performance on the on-time delivery measure, but, as described below, have differential performance on the other four measures. Of these four measures, two are objective (warehouse capability utilization and equipment maintenance and adjustment) because they are determined using data or records supplied by the company's information systems. The other two measures are subjective because they are rated by the company's regional vice president of operations, who directly supervises the two distribution centers. Importantly, participants are explicitly told that 'Past experience suggests that the accuracy and reliability of this vice president's rating in reflecting actual performance is approximately the same as those of the measures determined using data or records from the company's information systems.'[3] For each measure, participants are provided with: (1) a brief explanation of the measure, (2) how the measure is determined, (3) a performance target set by the company, (4) the manager's actual performance, and (5) the percentage of actual performance in excess of the target. After reviewing the performance information, participants issue an overall performance rating for the manager.

In the experiment, the actual performance of the two managers (i.e. the manager of Distribution Center A and the manager of Distribution Center B; hereafter referred to as manager A and manager B) is always higher than the target.[4] Our experiment has a 1×2 design, in which we manipulate within-participant whether a manager performed better on subjective measures or objective measures as compared to the other manager. Specifically, as indicated earlier, the two managers have the same performance on the on-time delivery measure. For the other four measures (i.e. two subjective and two objective), the actual performance of one manager is on average 10.3% above the target on the two subjective measures and 4.9% above the target on the two objective measures. We refer to this manager as the SUBJ-BETTER type. By comparison, the actual performance of the other manager is on average 4.9% above the target on the two subjective measures and 10.3% above the target on the two objective measures. We refer to this manager as the OBJ-BETTER type.

Each participant is provided with performance information for both managers and is asked to issue an overall performance rating for each manager.[5] Table 1 presents a sample of the performance information provided to participants. As shown in Table 1, performance information for

[3]The information that subjective ratings are given by the two managers' direct supervisor and that these ratings are as accurate and reliable as other measures helps ensure that the rater is perceived as competent and unbiased. We chose not to directly comment on the rater's competence or independence in experimental instructions because (1) such comments might generate an experimenter demand effect leading participants to unduly favor the subjective measures, and (2) there is no obvious reason to believe that participants would systematically perceive the rater as incompetent or non-independent without such comments.

[4]We made both managers' actual performance higher than the target because we wanted to avoid any potential effect of "loss aversion" (Tversky & Kahneman, 1991) on participants' decision if actual performance were below the target.

[5]Following prior experimental studies (e.g., Banker et al., 2004; Libby et al., 2004; Lipe & Salterio, 2000), we use a within-participant design (i.e., each participant rates the performance of two managers) to control for participants' individual differences and increase statistical power (Charness & Kuhn, 2011). In the conclusion section, we note that this could make the contrast between objective and subjective measures more salient than if the managers are evaluated separately.

Table 1. A sample of the performance information provided to participants in Experiment One

Measures	Target	Actual performance		Percent better than target	
		Distribution Center A	Distribution Center B	Distribution Center A	Distribution Center B
On-time delivery (i.e. the percentage of orders delivered within contractual time limit) Source of data: Product delivery records and original orders/contracts	90%	95%	95%	5.56%	5.56%
Warehouse capability utilization (i.e. the percentage of warehouse capability used in operations) Source of data: Warehouse inspection records	82%	86.14%	90.63%	5.05%	10.52%
Equipment maintenance and adjustment (i.e. the percentage of warehousing and delivery equipment's timely maintenance as required by service standards and timely adjustment as needed by order processing) Source of data: Equipment management records	86%	90%	94.65%	4.65%	10.06%
Emergency and complexity management (i.e. the ability to resolve emergent or unusual complex issues and minimize the negative impact of such issues on the company) Source of data: Rated by the regional vice president (operation) on a scale from 0 to 100, with 100 being the best performance	80	88	84	10%	5%
Employee ability enhancement (i.e. the effectiveness in increasing employees' work-related knowledge, skills and capabilities) Source of data: Rated by the regional vice president (operation) on a scale from 0 to 100, with 100 being the best performance	85	94	89	10.59%	4.71%

manager A and manager B is presented, respectively, in two separate columns of the same spreadsheet. We randomly vary the order in which the SUBJ-BETTER type and the OBJ-BETTER type are presented to the participant: for some participants, manager A is the SUBJ-BETTER type and manager B is the OBJ-BETTER type; whereas for other participants, manager A is the OBJ-BETTER type and manager B is the SUBJ-BETTER type.

3.2. Participants and Experimental Procedures

Seventy-one professional managers participated in Experiment One. These participants were recruited from the Executive Development Program (EDP) at a major university in China. The EDP is a part-time non-degree program designed to provide business training for corporate managers. Students normally take classes on weekends and are granted a training certificate

upon the completion of a series of business courses. The instructor of one of the required EDP courses allowed us to conduct the study in his class. On the day when the study was conducted, this instructor introduced one of the coauthors to students near the end of the class. The coauthor briefly described the study to students as a business decision-making experiment and invited them to stay for about 20 minutes to participate (no monetary payment or course credit was offered for participation). Approximately 85% of the students agreed to participate. Subsequently, the coauthor administered the experiment in the classroom.

After the experiment began, instructions were distributed to participants. The instructions provided background information about the corporation and the two distribution centers and described the experimental task.[6] After reading the instructions, participants completed a quiz to ensure that they fully understood the experimental task. Then, participants were given information about each manager's performance (similar to the information shown in Table 1). This performance information is identical for all participants except that, as described earlier, whether manager A is the SUBJ-BETTER or OBJ-BETTER type is randomly varied between participants. After reviewing all information received, participants gave an overall performance rating for each manager on a 101-point scale where 0 = 'Reassign' and 100 = 'Excellent' (Lipe & Salterio, 2000). Finally, participants completed a post-experimental questionnaire.

Participants have an average of 6.3 years of full-time work experience, 3.7 years of supervisory experience, and 3.1 years of performance-evaluation experience. As shown in Panel A of Table 2, participants come from a variety of industries. More than two-thirds of the participants have a Bachelor's degree or higher. One quarter of the participants are top or senior managers (including the head of a company or a branch company and other types of senior-level corporate positions), and 72% are middle managers (including the head of a division or a functional team and other types of middle-level corporate positions). Thirty-four participants are randomly assigned to the group in which manager A is the SUBJ-BETTER type and 37 to the group in which manager B is the SUBJ-BETTER type. A cross-tabulation chi-squared test shows that the distribution of participants' academic degree, position, and industry does not differ significantly between the two groups.

3.3. Results

The mean and standard deviation of performance ratings for the SUBJ-BETTER and OBJ-BETTER types of managers are reported in the first two rows of Panel A of Table 3. We also break down the data based on the order in which the performance information of these two types is presented to participants (i.e. when manager A is the SUBJ-BETTER type and manager B is the OBJ-BETTER type, versus when manager B is the SUBJ-BETTER type and manager A is the OBJ-BETTER type). Our hypothesis predicts that the performance rating will be influenced more by objective measures than by subjective measures. In Experiment One, while the two managers outperformed each other on different measures by the same margin, we expect that participants will focus more on the OBJ-BETTER type's superior objective-measure performance than the SUBJ-BETTER type's superior subjective-measure performance. To test our prediction, we create a variable to proxy for the extent to which objective measures outweigh subjective measures in performance evaluation. This variable, labeled *OUTWEIGH*, equals the performance rating for the OBJ-BETTER manager minus the performance rating for the SUBJ-BETTER manager. A positive *OUTWEIGH* (i.e. participants give a higher rating for the OBJ-BETTER manager)

[6]Experimental instruments are originally written in English and translated to Chinese (i.e., participants' native language) using the back-translation method (Brislin, 1970, 1986).

Table 2. The distribution of participants' education level, position, and industry

Panel A: Experiment One

Education			Position			Industry		
	#	%		#	%		#	%
Doctorate	2	3%	Head of a company	5	7%	Manufacturing	6	8%
Master	14	20%	Head of a branch company	2	3%	Construction	6	8%
Bachelor	39	55%	Other senior-level positions	11	15%	Transportation	2	3%
Associate	14	20%				Merchandising	20	28%
Lower than Associate	1	1%	Head of a division	17	24%	Agriculture	1	2%
Others	1	1%	Head of a functional group	8	11%	Finance	14	20%
			Other middle-level positions	26	37%	Service	17	24%
						Others	5	7%
			Others	2	3%			
Total	71		Total	71		Total	71	

Panel B: Experiment Two

Education			Position			Industry		
	#	%		#	%		#	%
Doctorate	0	0%	Head of a company	2	5%	Manufacturing	11	27.5%
Master	3	7.5%	Head of a branch company	1	2.5%	Construction	4	10%
Bachelor	25	62.5%	Other senior-level positions	2	5%	Transportation	5	12.5%
Associate	10	25%				Merchandising	5	12.5%
Lower than Associate	1	2.5%	Head of a division	22	55%	Agriculture	1	2.5%
Others	1	2.5%	Head of a functional group	2	5%	Service	5	12.5%
			Other middle-level positions	7	17.5%	Others	9	22.5%
			Others	4	10%			
Total	40		Total	40		Total	40	

would suggest that participants are influenced more by objective measures than by subjective measures in performance evaluation.

The third row of Panel A of Table 3 reports the mean and standard deviation of the *OUTWEIGH* measure. We again break down the data by the order in which the performance information of the SUBJ-BETTER and OBJ-BETTER types is presented to participants. As reported in Panel B of Table 3, a one-sample *t*-test reveals that the mean *OUTWEIGH* (1.73) is significantly higher ($p = 0.012$) than zero. Further, an independent-samples *t*-test shows that *OUTWEIGH* does not vary by the order of performance information presentation ($p = .914$). These results provide support to our hypothesis.

Our theory suggests that placing more weight on objective measures than subjective measures is driven by individuals' heuristic that objective measures are more scientific than subjective measures. To shed more light on participants' decision-making processes, in the post-experimental questionnaire, participants indicate the extent to which they agree with two statements, 'My performance rating is scientific,' and 'My performance rating is objective,' on an 11-point scale (1 = 'completely disagree' and 11 = 'completely agree'). A factor analysis reveals that responses to these two items load on a single factor (eigenvalue = 1.61; factor loadings = 0.90; proportion explained = 81%). We use the average of the two responses (Cronbach's alpha = 0.76) as a proxy for the participant's perception of the scientific nature of his performance-evaluation decision (labeled *SCIENTIFIC*). As reported in Panel C of Table 3, a regression of *OUTWEIGH* on *SCIENTIFIC*, controlling for the order of performance information presentation, reveals a marginally significantly positive relation ($p = .0997$), which is consistent with our theory that participants give more weight to objective measures because they

Table 3. Hypothesis tests for Experiment One

	When Manager A is SUBJ-BETTER	When Manager A is OBJ-BETTER	Row Mean
Panel A: Descriptive statistics			
The SUBJ-BETTER manager (i.e. the manager who performed better on subjective measures but worse on objective measures)	81.47 [7.39] $N = 34$	80.14 [8.72] $N = 37$	80.77 [8.08] $N = 71$
The OBJ-BETTER manager (i.e. the manager who performed better on objective measures but worse on subjective measures)	83.12 [6.87] $N = 34$	81.95 [6.92] $N = 37$	82.51 [6.88] $N = 71$
The *OUTWEIGH* measure	1.65 [5.65] $N = 34$	1.81 [6.90] $N = 37$	1.73 [6.28] $N = 71$

Panel B: Main tests

	t-statistic	*p*-value
One-sample *t*-test: *OUTWEIGH* > 0	2.32	0.012*
Two-sample *t*-test: Whether *OUTWEIGH* differs by the order of performance information presentation	0.11	0.914

Panel C: Additional analysis

OLS regression (Dependent variable = *OUTWEIGH*)

	Coefficient	Std. Err.	t-statistic	*p* value
Intercept	−2.41	3.85	−0.63	0.532
SCIENTIFIC	0.64	0.49	1.30	0.0997*
Order	−0.31	1.54	−0.20	0.841

Notes: Panel A reports the mean, [standard deviation], and the number of observations for: (1) the performance rating, which is given on a 101-point scale where 0 = 'Reassign' and 100 = 'Excellent,' for the SUBJ-BETTER type and the OBJ-BETTER type (the first and second rows), and (2) the *OUTWEIGH* measure, which equals the performance rating for the OBJ-BETTER manager minus the performance rating for the SUBJ-BETTER manager (the third row).
SCIENTIFIC = the average of participants' ratings of the degree to which they agree with the statements 'My performance rating is scientific' and 'My performance rating is objective,' on an 11-point scale where 1 = 'completely disagree' and 11= 'completely agree.'
Order = the order in which the SUBJ-BETTER type and the OBJ-BETTER type are presented to the participant (i.e. whether manager A is SUBJ-BETTER or OBJ-BETTER).
An asterisk indicates a one-tailed *p* value for testing a directional prediction.

believe that it makes their decision more scientific. This result provides further support for our hypothesis.

4. Experiment Two

Consistent with our theory, the results of Experiment One suggest that individuals assign more weight to objective measures than subjective measures, even when they are provided with historical information suggesting that the measures are equally accurate and reliable. However, one could argue that the objective measures used in Experiment One (warehouse capability utilization

and equipment maintenance and adjustment) seem more specific and easier to understand than the subjective measures (emergency and complexity management and employee ability enhancement). Thus, participants may have relied more on information that they deem to be more precise or descriptive, and therefore may have been influenced by the specificity of the measures rather than the objective or subjective nature of the measures. To further examine the robustness of the results of Experiment One and give our theory a more direct test, we design a second experiment.

4.1. Experimental Setting and Design

The basic setting and task of Experiment Two are similar to those of Experiment One. The major difference is that, in Experiment Two, participants evaluate the two managers' performance based on three measures: on-time delivery, warehouse capability utilization (these two measures are determined the same way as in Experiment One), and innovation and improvement (detail provided below). The actual performance of manager A and manager B is the same on the on-time delivery measure. For the other two measures, manager A's actual performance is 10% above the target on the innovation and improvement measure and 5.26% above the target on the warehouse capability utilization measure, whereas manager B's actual performance is 5% above the target on the innovation and improvement measure and 10.52% above the target on the warehouse capability utilization measure. That is, manager A always performed better than manager B on the innovation and improvement measure.

The warehouse capacity utilization measure is always an objective measure. We manipulate between participants whether the innovation and improvement measure (i.e. the measure on which manager A performed better than manager B) is subjective or objective, resulting in a 1×2 design.[7] In one condition, the innovation and improvement measure is rated by the regional vice president of operations, who directly supervises the two managers. As in Experiment One, participants are told that past experience suggests that the accuracy and reliability of this vice president's rating in reflecting actual performance is approximately the same as those of the objective measures. We refer to this condition as the 'Innovation-subjective' condition. In the other condition, the innovation and improvement measure is determined as the growth rates of innovation and improvement projects, compiled based on relevant project reports and records. We refer to this condition as the 'Innovation-objective' condition. Each participant is provided with performance information for both managers and is asked to give an overall performance rating for each manager. Table 4 presents a sample of the performance information provided to participants.

4.2. Participants and Experimental Procedures

Forty professional managers participated in Experiment Two. Participants were recruited from a different cohort of EDP students at the same university as in Experiment One. Experiment Two was conducted in a way similar to Experiment One and a coauthor administered the experiment in an EDP class. Participants have an average of 16.9 years of full-time work experience, 9.2 years of supervisory experience, and 4.9 years of performance-evaluation experience. As shown in Panel B of Table 2, participants work in a variety of industries. Seventy percent of the participants have a Bachelor's degree or higher. One-eighth of the participants are top or senior managers, and nearly 80% are middle managers. Nineteen participants are randomly assigned to the Innovation-subjective condition and 21 to the Innovation-objective condition. A cross-tabulation chi-squared

[7]In Experiment Two, we do not vary whether it is manager A or manager B who performed better on the innovation and improvement measure because the results of Experiment One show that this order has no impact on participants' decision-making.

Table 4. A sample of the performance information provided to participants in Experiment Two

Measures	Target	Actual performance		Percent better than target	
		Distribution Center A	Distribution Center B	Distribution Center A	Distribution Center B
On-time delivery (i.e. the percentage of orders delivered within contractual time limit) Source of data: Product delivery records and original orders/contract	90%	95%	95%	5.56%	5.56%
Warehouse capability utilization (i.e. the percentage of warehouse capability used in operations) Source of data: Warehouse inspection records	82%	86.31%	90.63%	5.26%	10.52%
Innovation and improvement (i.e. the effectiveness in improving, optimizing, and innovating the current operating procedures) Source of data: Rated by the regional vice president (operation) on a scale from 0 to 100, with 100 being the best performance	80	88	84	10%	5%

test shows that the distribution of participants' academic degree, position, and industry does not significantly differ between these two conditions. The experimental procedures are the same as those in Experiment One, except for the differences in experimental instruments between the two experiments.

4.3. Results

The first two columns of Panel A of Table 5 report the mean and standard deviation of the performance rating in the two experimental conditions. Our hypothesis predicts that participants' performance ratings will be influenced more by objective measures than by subjective measures because they heuristically perceive objective measures to be more scientific than subjective measures. In the Innovation-objective condition, the two measures on which the two managers have differential performance are both objective, and there is no subjective measure. The heuristic described above plays no role in this setting and, thus, we do not expect performance ratings to differ between the two managers. We use this condition as a baseline. By contrast, in the Innovation-subjective condition, performance is evaluated based on both objective and subjective measures, and the lay-scientism heuristic now has an impact on decision-making. Because manager A performed worse on an objective measure (although better on a subjective measure) than manager B, we expect that participants will be more sensitive to the objective measure than the subjective measure and, therefore, give a lower performance rating for manager A than for manager B. That is, we expect performance ratings to be different between the two managers.

To test our prediction, we create a variable, which equals the rating for manager B minus the rating for manager A, to proxy for the extent to which objective measures outweigh subjective measures in decision-making (labeled *OUTWEIGH*). As discussed above, we expect *OUTWEIGH* to be positive in the Innovation-subjective condition but not significantly different from zero in the Innovation-objective condition. The third column of Panel A of Table 5 presents the mean and standard deviation of *OUTWEIGH* for the two experimental conditions.

Table 5. Hypothesis tests for Experiment Two

	Manager A's rating	Manager B's rating	The *OUTWEIGH* measure
Panel A: Descriptive statistics			
The Innovation-subjective condition (i.e. the innovation and improvement measure is subjective)	83.22 [7.51] $N = 19$	85.01 [5.11] $N = 19$	1.79 [5.44] $N = 19$
The Innovation-objective condition (i.e. the innovation and improvement measure is subjective)	80.19 [9.64] $N = 21$	78.29 [12.78] $N = 21$	− 1.90 [9.08] $N = 21$
		t-statistic	*p*-value
Panel B: Statistical tests			
Two-sample *t*-test: *OUTWEIGH* in the Innovation-subjective condition > *OUTWEIGH* in the Innovation-objective condition		1.54	0.066*
One-sample *t*-tests:			
OUTWEIGH > 0 in the Innovation-subjective condition		1.44	0.084*
OUTWEIGH = 0 in the Innovation-objective condition		− 0.96	0.348

Notes: Panel A reports the mean, [standard deviation], and the number of observations for performance rating (the first two columns), which is given on a 101-point scale where 0 = 'Reassign' and 100 = 'Excellent', and the *OUTWEIGH* measure (the third column), which equals the rating for manager B minus the rating for manager A, in the two experimental conditions.
An asterisk indicates a one-tailed *p* value for testing a directional prediction.

As reported in Panel B of Table 5, an independent-samples *t*-test shows that, consistent with our prediction, *OUTWEIGH* is marginally significantly higher ($p = .066$) in the Innovation-subjective condition (1.79) than in the Innovation-objective condition (-1.90). Furthermore, one-sample *t*-tests find that *OUTWEIGH* is marginally significantly higher ($p = .084$) than zero in the Innovation-subjective condition but is not significantly different ($p = .348$) from zero in the Innovation-objective condition. These results support our hypothesis.[8]

We use the same two post-experimental items as in Experiment One to proxy for participants' perceptions of the scientific nature of their decisions: that is, participants' agreement with the statements, 'My performance rating is scientific,' and 'My performance rating is objective' (1 = 'completely disagree' and 11 = 'completely agree'). A factor analysis shows that responses to these two items load on a single factor (eigenvalue = 1.71; factor loadings = 0.92; portion explained = 85%). We create a variable, labeled *SCIENTIFIC*, by taking the average of these two items (Cronbach's alpha = 0.83).

[8]In Experiment Two, the innovation and improvement measure is a rating given by the superior on a 100-point scale in the Innovation-subjective condition and is a percentage growth rate calculated based on relevant project reports in the Innovation-objective condition. To investigate whether this difference in measurement metrics affects our results, we re-run Experiment Two with one change: in the Innovation-subjective condition the innovation and improvement measure is the same growth rate as in the Innovation-objective condition but participants are told that it is estimated by the superior. Results show that, similar to Experiment Two, *OUTWEIGH* is significantly higher ($p = .049$, two-tailed) than zero in the Innovation-subjective condition but is not significantly different ($p = .122$, two-tailed) from zero in the Innovation-objective condition. Therefore, we find no evidence that the results of Experiment Two are influenced by measurement metrics, which lends credence to our theory.

As discussed earlier, our theory suggests that the lay-scientism heuristic plays a bigger role when performance evaluation is based on a mix of objective and subjective measures than when performance evaluation is based on objective measures only. Specifically, when all performance measures are objective (i.e. the Innovation-objective condition), the scientific nature of these measures is not the focus of attention because it cannot be used to differentiate the two managers' performance. So, participants may be relatively insensitive to whether their performance-evaluation decision is scientific. By comparison, the performance-evaluation decision seems more scientific when it is based on the perceived scientific nature of objective versus subjective measures (i.e. the Innovation-subjective condition). Further, the more strongly participants are influenced by the lay-scientism heuristic, the more weight they may give to objective measures relative to subjective measures. Therefore, empirically, we expect that the *SCIENTIFIC* measure will be higher in the Innovation-subjective condition than in the Innovation-objective condition and will drive the difference in performance ratings between the two managers in the Innovation-subjective condition.

We use the Baron and Kenny (1986) procedures to test whether *SCIENTIFIC* mediates the difference in *OUTWEIGH* between the Innovation-subjective and Innovation-objective conditions. As depicted in Figure 1, first, we find that *OUTWEIGH* is marginally significantly higher ($p = .066$) in the Innovation-subjective condition than in the Innovation-objective condition. Second, as predicted, *SCIENTIFIC* is significantly higher ($p = .053$) in the Innovation-subjective condition than in the Innovation-objective condition. Third, when *OUTWEIGH* is regressed on *SCIENTIFIC* and a condition variable, *SCIENTIFIC* is marginally significant ($p = .064$) whereas the condition variable becomes non-significant ($p = .424$).[9] Overall, these results suggest that, consistent with our theory, *SCIENTIFIC* fully mediates the difference in *OUTWEIGH* between the two experimental conditions.[10]

5. Alternative Explanations

The results of the two experiments are consistent with our theory, but they also are consistent with two alternative explanations. First, participants could give more weight to objective measures simply because the metrics used as objective measures in the experiments are considered more important for accurately assessing distribution center managers' performance.[11] Second, participants may give less weight to subjective measures because they believe that these measures should have been objectively determined. That is, participants may question the appropriateness of subjectively measuring certain performance dimensions. We conduct additional analyses to address these two issues.

---

[9]The significant correlation between *SCIENTIFIC* and *OUTWEIGH* is consistent with our theory that the mix of objective and subjective measures in the Innovation-subjective condition leads to higher *SCIENTIFIC*, which in turn leads to greater *OUTWEIGH*. However, it is possible that the higher *SCIENTIFIC* reflects participants' *ex post* perceptions about their differential weighting of measures rather than the mindset that drives such weighting. To help address this issue, we conduct a *t*-test to compare *SCIENTIFIC* between the two experimental conditions, after dropping participants who place more weight on objective measures (i.e., observations with *OUTWEIGH* > 0) in the Innovation-subjective condition. If our theory holds, we would expect that even participants who do *not* place more weight on objective measures in the Innovation-subjective condition will perceive their decisions as more scientific than participants in the Innovation-objective condition. Indeed, test results show that *SCIENTIFIC* is marginally significantly higher (one-tailed $p = .091$) for the former group than for the latter group, thus lending credence to our theory.

[10]We repeat the mediation test using a Sobel test with the z-prime adjustment (MacKinnon et al., 2002; Mallinckrodt, Abraham, Wei, & Russell, 2006), and statistical inferences are unchanged.

[11]For example, participants could believe that, for distribution center managers, "warehouse capability utilization" is a more important indicator of the "right" performance than "emergency and complexity management", regardless of whether they are objectively or subjectively measured.

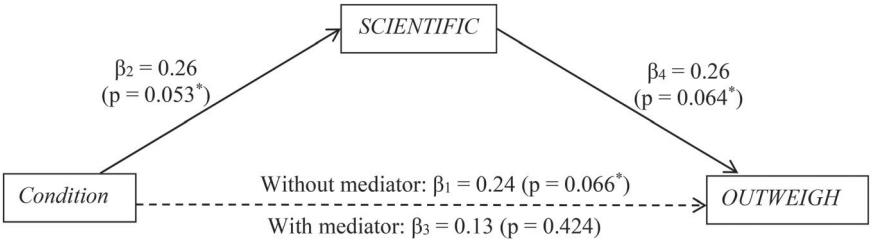

Figure 1. Mediation test for Experiment Two.
Notes: This figure reports the results of a test of the mediation effect of *SCIENTIFIC* on the difference in *OUTWEIGH* between the Innovation-subjective and Innovation-objective conditions, using the Baron and Kenny (1986) procedures. Path coefficients represent standardized regression weights (p values are reported in parentheses) from the following three models:

(1) $OUTWEIGH = \alpha + \beta_1\ Condition + \varepsilon$
(2) $SCIENTIFIC = \alpha + \beta_2\ Condition + \varepsilon$
(3) $OUTWEIGH = \alpha + \beta_3\ Condition + \beta_4\ SCIENTIFIC + \varepsilon$

SCIENTIFIC = the average of participants' ratings of the degree to which they agree with the statements 'My performance rating is scientific' and 'My performance rating is objective,' on an 11-point scale where 1 = 'completely disagree' and 11 = 'completely agree.'
OUTWEIGH = the rating for manager B – the rating for manager A.
Condition = one if the condition is Innovation-subjective and zero if the condition is Innovation-objective.
An asterisk indicates a one-tailed p value for testing a directional prediction.

5.1. Relative Importance of Performance Indicators

To investigate whether the perceived importance of performance indicators differs in our experiments, we recruited 34 managers from an EDP cohort different from those used in Experiments One and Two at the same university.[12] These managers have an average of 7.6 years of full-time work experience, 3.6 years of supervisory experience, and 3 years of performance-evaluation experience. We provided them with the same background information (i.e. information about XYZ Corporation and the two distribution centers) as used in Experiments One and Two. Then, we gave them a set of performance indicators, including the ones used in our experiments. For each performance indicator, we provided a brief explanation for the meaning of the metric but no information about whether the metric is objectively or subjectively measured.[13] We asked participants to indicate how important they thought each performance indicator was for accurately and reasonably evaluating the performance of distribution center managers, on an 11-point scale (1 = 'Not important at all' and 11 = 'Very important').

Of the five performance indicators on which manager A and manager B have differential performance in our two experiments, the mean importance rating is 8.3 for warehouse capability utilization, 8.3 for equipment maintenance and adjustment, 8.5 for emergency and complexity management, 8.2 for employee ability enhancement, and 8.2 for innovation and improvement. One-sample t-tests show that all five ratings are significantly higher (p values $< .001$) than the scale's midpoint of 6, suggesting that participants generally believe that these metrics are important for performance evaluation. A repeated-measures ANOVA reveals that the importance rating is not significantly different ($F = 0.53$, $p = .711$) across the five measures. Therefore,

[12]We did not collect the data needed to address alternative explanations directly from participants in Experiments One and Two because, in that case, such data could be contaminated by the decisions that participants made in the experiment.
[13]For example, we explain that "warehouse capability utilization" measures how efficiently the distribution center utilizes its available warehouse capability and "emergency and complexity management" measures the distribution center's ability to effectively resolve unpredictable, emergent problems and minimize the negative impact of such problems on the firm.

we find no evidence that different levels of importance are attached to performance indicators used in our experiments.

5.2. Appropriateness of Subjective Measurement

To investigate whether subjective measurement is considered inappropriate in our experimental setting, we recruit 40 managers (with an average of 8.4 years of work experience, 3.7 years of supervisory experience, and 3.7 years of performance-evaluation experience) from an EDP cohort that we have not used previously. We gave them the same background information as used in our experiments. We then provided them with a set of performance measures, including the ones used in our experiments, and asked them to indicate on an 11-point scale whether each measure should be determined using data supplied by the company's information systems (endpoint of 1) or using relevant personnel's judgment (endpoint of 11).

We first consider the four measures on which the two managers have differential performance in Experiment One. The mean rating is 3.6 for warehouse capability utilization and 4.3 for equipment maintenance and adjustment. These two ratings are not significantly different ($t = 0.97$, $p = .337$) from each other and both are significantly lower (p values $< .001$) than the scale's midpoint of 6, suggesting that participants generally believe that these two measures should be objectively determined. The mean rating is 7.4 for emergency and complexity management and 6.7 for employee ability enhancement. These two ratings are not significantly different ($t = 1.05$, $p = .299$) from each other and are significantly (or marginally significantly) higher (p values $< .067$) than the scale's midpoint of 6, suggesting that participants believe that these two measures should be determined based on human judgments. Hence, we find no evidence that the appropriateness of subjective measurement is questioned in Experiment One.

Then, we consider the innovation and improvement measure, which is manipulated as either subjective or objective in Experiment Two. The mean rating for this measure is 6.7, which is not significantly different ($t = 1.40$, $p = .17$) from the scale's midpoint of 6, suggesting that participants do not have a clear preference regarding whether it should be subjectively or objectively determined (if anything, the rating is directionally leaning toward the 'subjective' side, which would work against the alternative explanation). This result provides reassurance that our manipulation of the innovation and improvement measure as subjective or objective does not seem to cause participants to question the appropriateness of the measurement. In sum, the results of our supplemental analyses do not support either of the two alternative explanations discussed above, and, thereby, provide further support for our hypothesis.

6. Conclusion and Discussion

In this paper, we use two experiments to investigate whether managers' heuristic reasoning affects their weighting of objective versus subjective measures and, thereby, influences their performance-evaluation decisions. Consistent with psychological theory, our results suggest that participants are influenced more by objective measures than by subjective measures in decision-making, even when there is historical information suggesting that these two types of measures are equally accurate and reliable. We find that such behavior is driven by participants' perception that objective measures are more scientific than subjective measures. Supplemental analyses help eliminate the alternative explanation that these results arise because participants believe the objective measures used in our experiments are more important for performance evaluation or that participants believe the subjective measures used in our experiments should, instead, be objectively determined.

Some limitations of this paper could be addressed in future research. The subjective measures used in our experiments are the superior's performance ratings. While participants were told that the superior's performance rating is as accurate and reliable as the objective measures in reflecting actual performance, we do not empirically test the extent to which they internalized this information. Future research can explore whether evaluators' personal experience has any impact on their judgment about the reliability of subjective measures. Moreover, it would be interesting to investigate whether evaluators' performance-evaluation decision differs when they receive information about the reliability of subjective measures that is consistent versus inconsistent with their personal belief. Our experimental instructions did not give participants any explicit description about the superior's credibility or trustworthiness because doing so could cause experimenter demand effects and the superior's credibility is not a variable of interest in this study. Prior research suggests that employees' reaction to performance appraisals is influenced by the appraiser's credibility (Fedor, Eder, & Buckley, 1989; Langan-Fox, Waycott, Morizzi, & McDonald, 1998). Future studies could explore whether the perceived credibility of the superior plays a role in evaluators' decision-making.

Our supplemental analysis suggests that the metrics used in our experiments are considered equally important for performance evaluation, independent of whether they are objectively or subjectively measured. Future research can investigate whether subjective measurement influences the perceived importance of the metrics and whether such influences are moderated by important contextual factors such as task complexity and the valence of performance. In addition, while our supplemental analysis helps eliminate two alternative explanations, participants' weighting of performance measures could still be influenced by other factors, such as the congruity of the measure (i.e. whether maximizing the measure also maximizes the likelihood of achieving the principal's goals) (Datar, Kulp, & Lambert, 2001). Future research may investigate whether congruity interacts with subjectivity to affect performance-evaluation decisions. We used practicing managers in China as experimental participants. While we are not aware of any existing theory that suggests that the Chinese culture may interact with our independent variables to influence participants' behavior, we cannot completely rule out the potential influences of national culture on individuals' decision-making in our study. In our experiment, we ask each participant to rate the performance of both managers. This within-participant design controls for participants' individual differences and increases the power of statistical tests. However, it is more likely to sensitize participants to the contrast between objective and subjective measures than if the two managers are evaluated separately. Future research can examine whether evaluators' preferences for objective measures differ between absolute versus relative performance evaluation.

Despite these potential limitations, the findings of our study make important contributions to the research literature on the role of cognitive factors in performance evaluation and also have important implications for organizations' control practices. The issues discussed above provide ample avenues for future research.

Acknowledgements

We thank Victor Maas (associate editor), two anonymous reviewers, Jason Brown, Changling Chen, Bryan Church, Lan Guo, Khim Kelly, Theresa Libby, and workshop participants at Wilfrid Laurier University for helpful comments.

Funding

We are grateful for the support of the National Natural Science Foundation of China (71302011; 71572034) and the UIBE grant CXTD5-02.

References

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

Bailey, W. J., Hecht, G., & Towry, K. L. (2011). Dividing the pie: The influence of managerial discretion extent on bonus pool allocation. *Contemporary Accounting Research*, *28*(5), 1562–1584.

Baiman, S., & Rajan, M. V. (1995). The informational advantages of discretionary bonus schemes. *The Accounting Review*, *70*(4), 557–579.

Baker, G. (2000). The use of performance measures in incentive contracts. *American Economic Review*, *90*(2), 415–420.

Baker, G., Gibbons, R., & Murphy, K. J. (1994). Subjective performance measures in optimal incentive contracts. *Quaterly Journal of Economics*, *109*(4), 1125–1156.

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

Banker, R. D., Chang, H., & Pizzini, M. J. (2004). The balanced scorecard: Judgmental effects of performance measures linked to strategy. *The Accounting Review*, *79*(1), 1–23.

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

Barlett, G., Johnson, E., & Reckers, P. (2014). Accountability and role effects in balanced scorecard performance evaluations when strategy timeline is specified. *European Accounting Review*, *23*(1), 143–165.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research – conceptual, strategic, and statistical considerations, *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Beach, L. R., & Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. *Academy of Management Review*, *3*(3), 439–449.

Bol, J. C. (2008). Subjectivity in compensation contracting. *Journal of Accounting Literature*, *28*(1), 1–24.

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

Bol, J. C., Kramer, S., & Maas, V. S. (2016). How control system design affects performance evaluation compression: The role of information accuracy and outcome transparency. *Accounting, Organizations and Society*, *51*, 64–73.

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

Boudon, R. (1998). Limitations of rational choice theory. *American Journal of Sociology*, *104*(3), 817–828.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, *1*(3), 185–216.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137–164). Thousand Oaks, CA: Sage.

Brown, H. I. (1988). *Rationality*. New York, NY: Routledge.

Budde, J. (2007). Performance measure congruity and the balanced scorecard. *Journal of Accounting Research*, *45*(3), 515–539.

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

Charness, G., & Kuhn, P. (2011). Lab labor: What can labor economists learn from the lab? In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics*, *Volume 4* (pp. 229–330), Part A. Amsterdam, The Netherlands: Elsevier.

Dalrymple, D. (2003). Scientific knowledge as a global public good: Contributions to innovation and the economy. In J. M. Esanu & P. F. Uhlir (Eds.), *The role of scientific and technical data and information in the public domain* (pp. 35–51). Washington, DC: The National Academic Press.

Dane, E., & Pratt, M. G. (2007). Exploring intuition and its role in managerial decision making. *Academy of Management Review*, *32*(1), 33–54.

Datar, S. M., Kulp, S. C., & Lambert, R. A. (2001). Balancing performance measures. *Journal of Accounting Research*, *39*(1), 75–92.

Du, Y., Deloof, M., & Jorissen, A. (2013). Headquarters-subsidiary interdependencies and the design of performance evaluation and reward systems in multinational enterprises. *European Accounting Review*, *22*(2), 391–424.

Fedor, D. B., Eder, R. W., & Buckley, M. R. (1989). The contributory effects of supervisor intentions on subordinate feedback responses. *Organizational Behavior and Human Decision Processes*, *44*, 396–414.

Feltham, G. A., & Xie, J. (1994). Performance measure congruity and diversity in multi-task principal/agent relations. *The Accounting Review*, *69*(3), 429–453.

Ford, J. K., Kraiger, K., & Schechtman, S. L. (1986). Study of race effects in objective indices and subjective evaluations of performance: A meta-analysis of performance criteria. *Psychological Bulletin*, *99*(3), 330–337.

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

Hayes, R. M., & Schaefer, S. (2000). Implicit contracts and the explanatory power of top executive compensation for future performance. *The Rand Journal of Economics*, *31*(2), 273–293.

Healy, P. (1993). Rationality, judgment, and critical inquiry. *Electronic Journal of Analytic Philosophy*, *1*(3).

Hoffman, C. C., Nathan, B. R., & Holden, L. M. (1991). A comparison of validation criteria: Objective versus subjective performance measures and self- versus supervisor ratings. *Personnel Psychology*, *44*, 601–618.

Hogarth, R. M., & Kunreuther, H. (1995). Decision making under ignorance: Arguing with yourself. *Journal of Risk and Uncertainty*, *10*, 15–36.

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics and Organizations*, *7*, 24–52.

Hoyer, W. D. (1984). An examination of consumer decision making for a common repeat purchase product. *Journal of Consumer Research*, *11*(3), 822–829.

Hsee, C. K., & Hastie, R. (2006). Decision and experience: Why don't we choose what makes us happy? *Trends in Cognitive Sciences*, *10*(1), 31–37.

Hsee, C. K., Zhang, J., Yu, F., & Xi, Y. (2003). Lay rationalism and inconsistency between predicted experience and decision. *Journal of Behavioral Decision Making*, *16*(4), 257–272.

Ireland, R. D., Hitt, M. A., Bettis, R. A., & Auld de Porras, D. (1987). Strategy formulation processes: Differences in perceptions of strength and weaknesses indicators and environmental uncertainty by managerial level. *Strategic Management Journal*, *8*(5), 469–485.

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., & Meyer, M. W. (2003). Subjectivity and the weighting of performance measures: Evidence from a balanced scorecard. *The Accounting Review*, *78*(3), 725–758.

Jarvis, R., Curran, J., Kitching, J., & Lightfoot, G. (2000). The use of quantitative and qualitative criteria in the measurement of performance in small firms. *Journal of Small Business and Enterprise Development*, *7*(2), 123–134.

Johnson, E. N., Reckers, P. M., & Barlett, G. D. (2014). Influences of timeline and perceived strategy effectiveness on balanced scorecard performance evaluation judgments. *Journal of Management Accounting Research*, *26*(1), 165–184.

Kaplan, R. S., & Norton, D. P. (1996). Using the balanced scorecard as a strategic management system. *Harvard Business Review*, *74*(1), 75–85.

Kiesler, S., & Sproull, L. (1982). Managerial response to changing environments: Perspectives on problem sensing from social cognition. *Administrative Science Quarterly*, *27*(4), 548–570.

Kramer, S., & Maas, V. (2016). *Selective attention to performance measures and bias in subjective performance evaluations: an eye-tracking study*. Working paper. Erasmus University of Rotterdam and University of Amsterdam.

Krishnan, R., Luft, J. L., & Shields, M. D. (2005). Effects of accounting-method choices on subjective performance-measure weighting decisions: Experimental evidence on precision and error covariance. *The Accounting Review*, *80*(4), 1163–1192.

Langan-Fox, J., Waycott, J., Morizzi, M., & McDonald, L. (1998). Predictors of participation in performance appraisal: A voluntary system in a blue-collar work environment. *International Journal of Selection and Assessment*, *6*(4), 249–260.

Lawler, E. E. (1971). *Pay and organizational effectiveness: A psychological view*. New York, NY: McGraw-Hill.

Libby, T., Salterio, S. E., & Webb, A. (2004). The balanced scorecard: The effects of assurance and process accountability on managerial judgment. *The Accounting Review*, *79*(4), 1075–1094.

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

Long, J. H., Mertins, L., & Vansant, B. (2015). The effects of firm-provided measure weightings on evaluators' incorporation of non-contractible information. *Journal of Management Accounting Research*, *27*(1), 47–62.

Luft, J. L. (2009). Nonfinancial information and accounting: A reconsideration of benefits and challenges. *Accounting Horizons*, *23*(3), 307–325.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to mediation and other intervening variable effects. *Psychological Methods*, *7*(1), 83–104.

Mallinckrodt, B., Abraham, W. T., Wei, M., & Russell, D. W. (2006). Advances in testing the statistical significance of mediation effects. *Journal of Counseling Psychology*, *53*(3), 372–378.

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

Muckler, F. A., & Seven, S. A. (1992). Selecting performance measures: "Objective" versus "subjective" measurement. *Human Factors*, *34*(4), 441–455.

Murphy, K. J. (1992). Performance measurement and appraisal: Motivating managers to identify and reward performance. In W. J. Bruns, Jr. (Eds.), *Performance measurement, evaluation, and incentives* (pp. 37–62). Boston, MA: Harvard Business School Press.

Murphy, K. J., & Oyer, P. (2003). *Discretion in executive incentive contracts: Theory and evidence*. Working paper. University of Southern California and Stanford University.

Nathan, B. R., & Alexander, R. A. (1988). A comparison of criteria for test validation: A meta-analytic investigation. *Personnel Psychology*, *41*, 517–535.

Ouellette, J. A., & Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin*, *124*(1), 54–74.

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

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

Rajan, M. V., & Reichelstein, S. (2009). Objective versus subjective indicators of managerial performance. *The Accounting Review*, *84*(1), 209–237.

Rich, G. A., Bommer, W. H., MacKenzie, S. B., Podsakoff, P. M., & Johnson, J. L. (1999). Apples and apples or apples and oranges? A meta-analysis of objective and subjective measures of salesperson performance. *Journal of Personal Selling and Sales Management*, *65*(4), 41–52.

Sadler-Smith, E. (2004). Cognitive style and the management of small and medium-sized enterprises. *Organization Studies*, *25*(2), 155–181.

Schwenk, C. R. (1988). The cognitive perspective on strategic decision making. *Journal of Management Studies*, *25*(1), 41–55.

Simon, H. A. (1987). Making management decisions: The role of intuition and emotion. *Academy of Management Executive*, *1*(1), 57–64.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology*, *28*, 1–39.

Tsai, C. I., & Hsee, C. K. (2009). A behavioral account of compensation awarding decisions. *Journal of Behavioral Decision Making*, *22*(2), 138–152.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference dependent model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061.

Van der Stede, W. A., Chow, C. W., & Lin, T. W. (2006). Strategy, choice of performance measures, and performance. *Behavioral Research in Accounting*, *18*, 185–205.

Wall, T. D., Michie, J., Patterson, M., Wood, S. J., Sheehan, M., Clegg, C. W., & West, M. (2004). On the validity of subjective measures of company performance. *Personnel Psychology*, *57*, 95–118.

Walsh, J. P. (1995). Managerial and organizational cognition: Notes from a trip down memory lane. *Organization Science*, *6*(3), 280–321.

Woods, A. (2012). Subjective adjustments to objective performance measures: The influence of prior performance. *Accounting, Organizations and Society*, *37*(6), 403–425.