THE ACCOUNTING REVIEW
Vol. 78, No. 3
2003
pp. 725–758

Subjectivity and the Weighting of Performance Measures: Evidence from a Balanced Scorecard

Christopher D. Ittner
David F. Larcker
Marshall W. Meyer
University of Pennsylvania

ABSTRACT: This study examines how different types of performance measures were weighted in a subjective balanced scorecard bonus plan adopted by a major financial services firm. Drawing upon economic and psychological studies on performance evaluation and compensation criteria, we develop hypotheses regarding the weights placed on different types of measures. We find that the subjectivity in the scorecard plan allowed superiors to reduce the "balance" in bonus awards by placing most of the weight on financial measures, to incorporate factors other than the scorecard measures in performance evaluations, to change evaluation criteria from quarter to quarter, to ignore measures that were predictive of future financial performance, and to weight measures that were not predictive of desired results. This evidence suggests that psychology-based explanations may be equally or more relevant than economics-based explanations in explaining the firm's measurement practices. The high level of subjectivity in the balanced scorecard plan led many branch managers to complain about favoritism in bonus awards and uncertainty in the criteria being used to determine rewards. The system ultimately was abandoned in favor of a formulaic bonus plan based solely on revenues.

Keywords: *balanced scorecard; subjective performance measures; nonfinancial performance measurement.*

Data Availability: *All data are proprietary. Confidentiality agreements prevent the authors from distributing the data.*

I. INTRODUCTION

This study investigates the use of subjectivity in reward systems containing multiple performance measures. Bonuses based solely on profits and other financial accounting numbers have been criticized for encouraging managers to sacrifice long-run

We appreciate the comments of the reviewer and seminar participants at the University of Alberta, University of Cambridge, Citibank Behavioral Science Research Council Conference, Georgia State University, University of New South Wales, University of Saskatchewan, Stanford University Summer Accounting Research Camp, The University of Texas at Austin, The University of Texas at Dallas, and University of Waterloo. This research was funded by the Citibank Behavioral Science Research Council, whose support is gratefully acknowledged.

Editor's note: This paper was accepted by Terry Shevlin, Senior Editor.

Submitted September 2002
Accepted February 2003

performance to increase short-term financial results, and thereby maximize their bonuses. To overcome the short-run orientation of accounting-based reward systems, many firms are implementing compensation plans that supplement financial metrics with additional measures in order to assess performance dimensions that are not captured in short-term financial results. These additional measures can take a variety of forms, ranging from quantitative, nonfinancial metrics, such as employee and customer survey results, to qualitative assessments of performance by the manager's superior.

One critical implementation issue that arises when incorporating multiple performance measures in reward systems is determining the relative weights to place on the various measures when determining bonuses. One option is to use a formula that explicitly weights each measure. Potential difficulties with this option include determining the appropriate weights to place on each measure, the ''game-playing'' associated with any formula-based plan, the possibility that bonuses will be paid even when performance is ''unbalanced'' (i.e., overachievement on some objectives and underachievement on others), and the likelihood that all relevant dimensions of managerial performance are not captured by the selected performance measures (e.g., Holmstrom and Milgrom 1991; Baker et al. 1994; Kaplan and Norton 1996).

A second option is to introduce subjectivity into the bonus award process. This subjectivity can take the form of flexibility in weighting quantitative performance measures when computing a manager's bonus, the use of qualitative performance evaluations, and/or the discretion to adjust bonus awards based on factors other than the measures specified in the bonus contract. Some theoretical work indicates that greater subjectivity can improve incentive contracting because it allows the firm to exploit noncontractible information that might otherwise be ignored in formula-based contracts, and to mitigate distortions in managerial effort by ''backing out'' dysfunctional behavior induced by incomplete objective performance measures (Baker et al. 1994; Baiman and Rajan 1995). However, other research suggests that subjectivity in reward systems can lower managers' motivation by allowing evaluators to ignore certain types of performance measures that are included in the bonus plan, permitting bonus payout criteria to change each period, and introducing favoritism and bias into the reward system (e.g., Prendergast and Topel 1993). As a result, managers will be less able to distinguish what constitutes good performance, less likely to believe that rewards are contingent on performance, and less convinced that performance criteria are being applied consistently across the organization.

Drawing upon economic and psychological studies on the choice of performance measures for performance evaluation and compensation purposes, we develop exploratory hypotheses regarding the weights placed on different types of performance measures (e.g., financial versus nonfinancial, quantitative versus qualitative, and input versus outcome) in subjective bonus computations. We test these hypotheses using quantitative and qualitative data gathered during an extensive, multiyear field investigation of a balanced scorecard bonus plan in the U.S. retail banking operations of Global Financial Services (GFS), a leading international financial services provider.

In 1993, GFS replaced a profit-based bonus plan for branch managers with a *formula-based* system that rewarded multiple accounting and growth measures once customer satisfaction and operational audit hurdles were achieved. This system changed rapidly during the nine quarters it was in use as the bank sought to eliminate gaming and promote performance across a broader set of measures. The formula-based plan was replaced in the second quarter of 1995 by a ''balanced scorecard'' bonus plan containing six categories of financial and nonfinancial performance measures, some of which were based on qualitative evaluations by the managers' supervisors. Unlike the earlier formula-based plan, *subjective*

weightings were used to aggregate the various scorecard measures when determining overall performance evaluations and bonuses.

Although Kaplan and Norton (2001) cite the GFS bonus plan as an example of a scorecard-based reward system that prevented managers from underperforming on any of the scorecard dimensions, we find that the use of subjectivity in weighting the scorecard measures allowed supervisors to ignore many performance measures, even though some of these measures were leading indicators of the bank's two strategic objectives (financial performance and growth in customers). Instead, short-term financial performance measures become the primary determinant of bonuses. In addition, a large proportion of branch managers' performance evaluations was based on factors *other than* the scorecard measures, even though discretion to consider other factors was not a component of the bonus plan. The move from the formula-based system to the more subjective scorecard led many branch managers to complain about favoritism in bonus awards and uncertainty in the criteria being used to determine rewards, and caused corporate executives and human resource managers to question the scorecard's use for compensation purposes. Ultimately, the company abandoned the scorecard plan at the end of 1998 in favor of a commission-style system based on revenues.

Our study makes four contributions to the performance evaluation and compensation literatures. First, we extend cross-sectional studies on the use of subjectivity and discretion in bonus plans (e.g., Murphy and Oyer 2001; Gibbs et al. 2002). Whereas these studies focus on the factors explaining *who* includes subjectivity in compensation contracts, our study emphasizes *how* subjectivity is incorporated into performance evaluations and bonus awards. Second, we provide further evidence on the influence of informativeness on performance measure weighting. Prior studies on the relative weights placed on financial and nonfinancial performance measures (e.g., Bushman et al. 1996; Ittner et al. 1997) generally include proxies for the noise in financial measures, but do not include direct measures of the informativeness of *nonfinancial* measures due to data constraints. The detailed data in our study allow us to develop stronger tests of economic theories that the relative weights placed on performance measures other than financial results are a function of their informativeness. Third, our research complements psychology-based experimental work on the importance placed on various types of performance measures by examining whether their experimental results hold in an actual performance-evaluation setting. Finally, although the ability to generalize our results is limited by the analysis of only a single firm, we provide one of the first detailed studies of scorecard-based compensation plans.[1] Despite survey evidence that a growing number of firms are using balanced scorecards for compensation purposes (Kaplan and Norton 2001, Chapter 10), relatively little is known about the implementation issues associated with scorecard-based reward systems.

The remainder of the paper is organized in five sections. Section II reviews related research on subjective performance evaluation and performance measure weighting, and develops our exploratory hypotheses. The research setting for our study is described in Section III. The statistical tests for our research hypotheses are presented in Section IV. In Section V we provide detailed qualitative analysis regarding the success of the balanced scorecard implementation. The final section summarizes our research results and discusses implications for future research.

[1] See Malina and Selto (2001) and Campbell et al. (2002) for field-based studies examining other uses of the balanced scorecard. Also see Banker et al. (2000) for field evidence on the implementation of a compensation plan incorporating nonfinancial measures.

II. LITERATURE REVIEW AND HYPOTHESES

Performance Measure Use in Subjective Performance Evaluation

Kaplan and Norton (1996, 10) argue that balanced scorecards should reflect four types of measures: (1) financial and nonfinancial; (2) external (financial and customer) and internal (critical business processes, innovation, and learning and growth); (3) inputs/drivers and outcomes/results; and (4) objective, easily quantifiable measures and more subjective, judgmental measures. Although Kaplan and Norton (1996, 2001) provide little guidance on how to combine or ''balance'' these disparate measures when evaluating managerial performance, they conjecture that the balanced scorecard renders subjective reward systems ''easier and more defensible to administer...and also less susceptible to game playing.'' (See Kaplan and Norton 1996, 220.)

Analytical research on the use of subjectivity in performance evaluation and compensation focuses on the benefits of subjective bonus awards (e.g., Baker et al. 1994; Baiman and Rajan 1995), the drawbacks of subjective performance evaluations (e.g., Prendergast and Topel 1996; MacLeod 2001), and the factors influencing the relative weights placed on subjective versus objective performance measures in incentive contracts (e.g., Murphy and Oyer 2001). Most of these models do not examine how different types of performance measures or different forms of subjectivity (i.e., flexibility in assigning weights to measures, use of qualitative performance evaluations, and/or discretion to incorporate other performance criteria) should be incorporated into subjective bonus awards. An exception is Murphy and Oyer's (2001) model, which suggests that the relative weight on subjective measures will be higher in privately held companies, larger companies with more top managers, less autonomous business units, companies with substantial growth opportunities, and companies where accounting profits and shareholder returns are less highly correlated. Their cross-sectional empirical tests of executive bonuses provide mixed support for these hypotheses.

A related empirical study of automobile dealerships by Gibbs et al. (2002) finds that subjectivity (defined as the presence of any subjective bonus payout or as ''discretionary bonus'' as a percent of total compensation) is positively related to departmental interdependencies, financial losses, the manager's tenure, and the achievability of formula-based bonuses. While these two empirical studies provide insight into *who* uses subjectivity in compensation contracts, they provide little insight into *how* subjectivity is applied or performance is evaluated when multiple types of performance measures are incorporated into the bonus contract.[2]

Other studies address the relative importance placed on the various types of measures highlighted by Kaplan and Norton (1996). These studies fall into two research streams. The first stream focuses on economic models of incentive contracting. Economics-based agency models emphasize the role of performance measure choice in aligning agents' goals with those of the principal, and indicate that the choice of performance measures in incentive contracts should be a function of the informativeness (or incremental information content) of each measure regarding the worker's action choices (e.g., Holmstrom 1979; Banker and Datar 1989; Feltham and Xie 1994; Hemmer 1996; Lambert 2001).

A second research stream adopts a psychological perspective. These studies examine how human information-processing capabilities and decision strategies influence the types of information individuals use when assessing performance. These behavioral experiments

[2] Murphy and Oyer's (2001) tests examine the use of individual performance appraisals and discretionary allocation of bonus awards, but do not examine the use of specific types of performance measures. Gibbs et al. (2002) use an aggregate measure of subjectivity, which their survey defines as bonuses awarded based on a supervisor's subjective judgment of a manager's performance.

suggest that issues such as information overload and cognitive biases can play a significant role in the relative weights placed on different types of balanced scorecard measures (e.g., Lipe and Salterio 2000, 2002). In particular, this research finds that evaluators frequently place greater or exclusive emphasis on certain types of measures, even when other types of measures also provide relevant information on the subordinate's performance.

Economics-Based Hypotheses

Informativeness

Economics-based agency models focus on the role of performance measures in promoting congruence between the principal's objective and that of the agent (e.g., Holmstrom 1979; Banker and Datar 1989; Lambert 2001). Two primary insights from these models are that compensation contracts should include any (costless) measure that carries incremental information on the agent's actions, and that the relative weight placed on an individual measure should be a function of the measure's signal-to-noise ratio, as reflected in its sensitivity (or the change in its mean value in response to a change in the agent's action) and precision (or the inverse of the variance in the measure given the agent's action).

Although many of these models say little about the *specific* types of performance measures that should be included in compensation contracts, several studies extend these papers to investigate the role of nonfinancial measures (Feltham and Xie 1994; Hauser et al. 1994; Hemmer 1996). These models suggest that financial measures alone are unlikely to be the most efficient means to motivate employees, and demonstrate how incentives based on nonfinancial measures can improve contracting by incorporating information on agents' actions that is not fully captured in contemporaneous financial results.

A number of cross-sectional empirical studies draw upon informativeness theories when examining the relative weights placed on individual, nonfinancial, or subjective performance measures (e.g., Bushman et al. 1996; Ittner et al. 1997; Murphy and Oyer 2001; Ittner and Larcker 2002). These studies typically use two approaches to test the models' predictions. First, the effects of noise on performance measure choice are examined using the variance in objective, financial measures. In these tests, the weight placed on financial measures is predicted to decrease in their noisiness, while the weight on other types of measures is predicted to increase. Second, proxies for factors that are expected to influence the informativeness of individual, nonfinancial, or subjective measures (e.g., growth opportunities, strategy, product life cycle, etc.) are examined, with the weight placed on these measures expected to be higher when the informativeness proxies are greater.

The results from these studies are mixed. While proxies for the factors predicted to influence informativeness are generally associated with increased weights on these measures, proxies for the noise in financial measures tend to have little association with measurement choices. However, a significant limitation of these analyses is the lack of data on nonfinancial or subjective performance dimensions, forcing researchers to use indirect proxies for the measures' informativeness.

Consistent with the nonfinancial performance measure and balanced scorecard literatures, we assume that current financial measures are potentially incomplete, and that other indicators of future financial performance can provide incremental information on the manager's actions (e.g., Feltham and Xie 1994; Hemmer 1996; Kaplan and Norton 1996, 2001). We also assume that measures that are more predictive of future performance provide greater information on the congruence between the agent's actions and the outcomes desired by the principal. In an agency setting, the coefficients (or weights) associated with the nonfinancial performance measures in the structural model linking nonfinancial measures to future financial results (i.e., the ''business model'') and the coefficients (or weights) used

in the agent's compensation contract will be identical if the agent is *risk neutral* (e.g., Datar et al. 2001). However, theoretical models by Gjesdal (1981) and Datar et al. (2001) also indicate that when the agent is *risk averse*, the coefficients in the business model will not be identical to the coefficients in the compensation model, complicating our hypothesis development and subsequent empirical tests.

Although the latter result is a potential concern, additional factors suggest that the statistical associations between nonfinancial measures and subsequent financial results are likely to provide reasonable proxies for the nonfinancial measures' informativeness in our setting. First, agency theory suggests that it would be unusual to observe a nonfinancial performance measure with a substantial positive coefficient in the business model (i.e., a measure that is highly predictive of future improvements in financial performance), but a zero or negative coefficient in the compensation model. This would only occur if the nonfinancial performance measure ''corrects'' the measurement error in another performance variable included in the model (i.e., a type of relative performance measurement), or if the use of the nonfinancial measure imposes too much risk on a risk-averse agent. We do not believe that our research setting exhibits either of these attributes.

Second, the lack of direct correspondence between the weights in the business model and the compensation model only arises in models that do not allow for private information by the agent. Once private information is incorporated into the formal agency model, the coefficients in the business model become quite similar (but not identical) to the coefficients in the compensation function. Since branch managers are likely to have private information about their operations and marketplace that is not available to their superiors, the comparison of weights between the business and compensation models should provide a reasonable test of informativeness theories. More specifically, if informativeness is a significant determinant of performance measurement choices, then we expect greater weight on nonfinancial measures that are more highly associated with future financial performance.

Subjective versus Objective Measures

Prendergast and Topel's (1996) principal-agent model extends these studies to examine how favoritism and bias in subjective performance evaluations affect the relative weights placed on objective versus subjective performance measures in compensation and promotion decisions. They argue that subjectivity opens the door to favoritism, where supervisors act on personal preferences toward subordinates to favor some employees over others. Their model indicates that favoritism leads firms to place too little weight on supervisor appraisals and other subjective opinions of performance and too much weight on ''hard'' performance measures when combining multiple indicators in order to constrain favoritism and reduce the noise in the performance information. These results suggest that greater weight will be placed on more objective, quantitative measures than on more subjective, qualitative measures.

Psychology-Based Hypotheses

Driver versus Outcome Measures

Psychology-based studies diverge from the rational choice models in the economics literature to investigate how human information-processing limitations and decision strategies influence the use of performance measures. One of the most frequently observed results relates to the ''outcome effect,'' in which evaluators systematically overweight outcome knowledge when assessing a manager's performance (e.g., Mitchell and Kalb 1981; Baron and Hershey 1988; Hawkins and Hastie 1990; Lipe 1993; Ghosh and Lusch 2000).

These studies find that evaluators tend to evaluate managers positively (negatively) when the outcome is positive (negative), regardless of whether the actions taken to achieve the results were appropriate. According to this literature, the overemphasis on outcomes arises because outcome knowledge influences the evidence recalled by the evaluator when attempting to assess the performance of the manager (Slovic and Fischhoff 1977). Consistent with these studies, we expect greater weight to be placed on outcome measures than on measures of the drivers of desired results, even when the driver measures are informative of the manager's actions.

In a similar vein, the balanced scorecard literature generally assumes that internal measures, such as innovation, process improvement, and employee satisfaction, are leading indicators or drivers of the outcome-oriented external (financial and customer) measures. If evaluators perceive external measures to be indicators of desired outcomes, then the outcome-effects literature suggests greater weight will be placed on external measures than on internal measures.

Financial versus Nonfinancial Measures

Relatively few psychology-based performance evaluation studies directly examine the relative weights placed on financial and nonfinancial measures for compensation purposes. Moreover, the results from related experimental studies provide mixed hypotheses regarding the relative weights placed on these measures. As discussed above, experimental evidence on outcome effects indicates that outcome or results-oriented measures will be weighted more heavily than input or driver measures. Since improved financial results are the ultimate goal of balanced scorecard systems (Kaplan and Norton 1996), outcome-effect studies also suggest that financial results will be weighted more heavily than nonfinancial results.

These studies also find that outcome effects are more pronounced when evaluators have used the outcome measures in earlier reward systems. Frederickson et al. (1999), for example, examine whether previous experience under a performance evaluation system systematically biases subsequent evaluations. They find that prior experience with an outcome-based performance evaluation system leads to larger outcome effects, with increased outcome feedback frequency increasing the effect. These results suggest that in a research setting such as ours, where financial results traditionally have been the primary performance objectives in bonus plans, greater weight will be placed on financial measures.

Lipe and Salterio (2000), in turn, examine whether evaluations using a balanced scorecard are affected by measures that are unique to a given organization, as well as measures that are common across organizations. Their experiment indicates that only common measures impact superior's evaluations, suggesting that measures that are used throughout an organization or are based on a common methodology will receive greater weight. To the extent that financial measures are more common and standardized across the organization's subunits, we expect greater weight on these measures.

More direct experiments on the use of financial and nonfinancial measures are inconclusive. Schiff and Hoffman (1996) find that executives tend to place greater weight on financial information when evaluating the performance of a business unit. However, they also find that participants place greater weight on nonfinancial information when evaluating a manager's performance. The authors provide no theoretical explanation for these differences.

Finally, Luft and Shields (2001) examine the use of financial and nonfinancial measures in a decision-making context. Their experiment indicates that participants place greater weight on current nonfinancial information when forecasting future financial performance

than on current financial information. The authors attribute this result to nonfinancial measures being more cognitively valuable (i.e., more meaningful, transparent, and understandable) than financial measures. However, their experiment does not address the use of financial and nonfinancial measures for performance evaluation. The conflicting results in the preceding studies make hypotheses regarding the relative weights on financial and nonfinancial measures in bonus decisions unclear.

Objective/Quantitative Measures versus Subjective/Qualitative Measures

The organizational psychology literature has long held that greater weight should be placed on performance measures that are more reliable (e.g., Bellows 1954; Blum and Naylor 1968). According to this literature, subjective, qualitative performance assessments are often less accurate and reliable than more objective, quantitative measures because they are influenced by the rater's biases (e.g., Feldman 1981; Heneman 1986; Campbell 1990). Support for claims that subjective measures are less reliable than objective, quantitative measures is provided by studies examining the associations between subjective and quantitative measures covering the same activities. Meta-analyses by Heneman (1986) and Bommer et al. (1995) find correlations of only 0.27 and 0.39, respectively, between objective and subjective performance evaluations. Their analyses (as well as Heneman et al.'s [1987] review of related experimental studies) also indicate that differences between subjective and objective measures can be substantially reduced when the subjective ratings are based on the aggregation of multiple measures rather than on a single overall rating, and when performance is rated relative to a target or peer group rather than on an absolute scale. Assuming reliability is an important factor in the choice of performance measures, these studies suggest that greater weight will be placed on quantitative measures than on qualitative measures, and that greater weight will be placed on measures that are based on aggregations of multiple indicators and performance relative to targets than on other measures.

III. RESEARCH SETTING

A summary of our research hypotheses is presented in Table 1. We test these hypotheses using data from the North American retail banking operations of Global Financial Services (GFS), a leading international financial services provider. Prior to the 1990s, GFS had a strong financial orientation that focused almost exclusively on the achievement of financial results. Following a significant downturn in financial performance in the early 1990s, the bank initiated a review to evaluate its strategic direction. One conclusion of this review was that the bank's problems could be traced to a single-minded focus on bottom-line earnings and revenue growth to the exclusion of other important issues. This conclusion led to the development of a new retail banking business model in 1992 (see Figure 1). According to this model, the traditional goals of market share growth and financial performance were driven by more fundamental factors including product and service quality, cost effectiveness, risk control, employee relations, innovation, and customer satisfaction. The bank's executives agreed that organizational success would require greater emphasis on *all* of these dimensions.

Performance Incentive Plan

To support the new business model, GFS modified its bonus plans to emphasize the perceived drivers of profitability and growth. Until 1992, GFS awarded bonuses to branch managers based on a single branch profitability measure. Beginning in 1993, this plan was

TABLE 1
Summary of Hypotheses regarding the Relative Weights Placed on Different Types of Performance Measures

Economics-Based Hypotheses

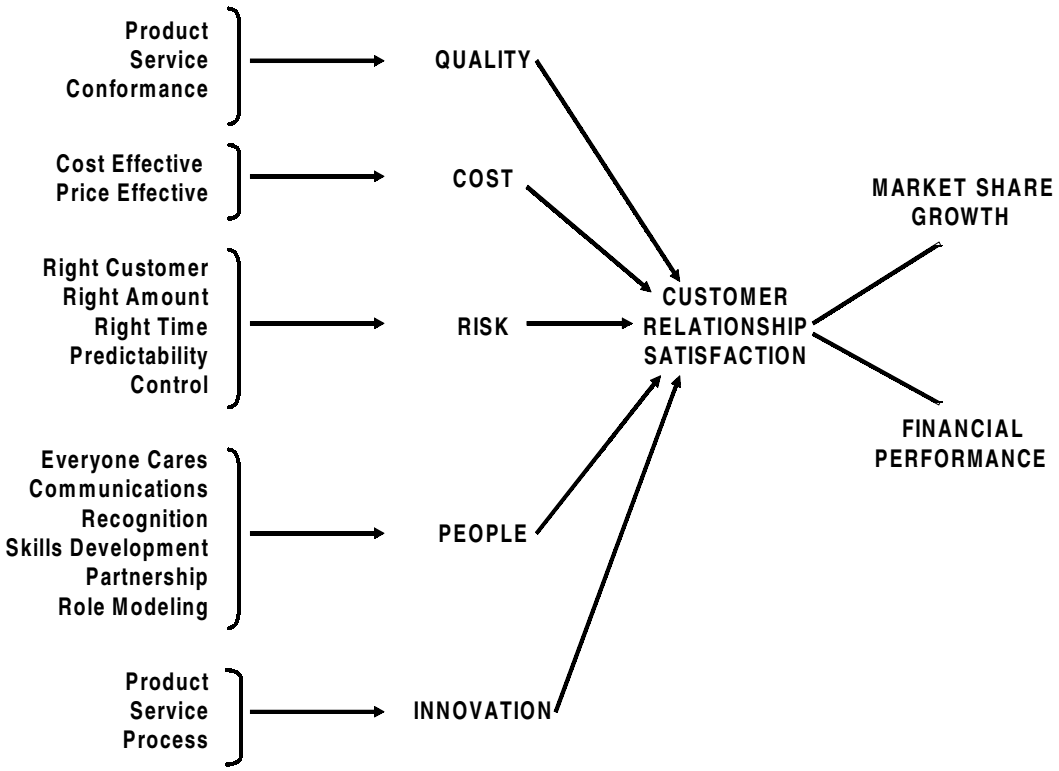
- Nonfinancial measures that are more predictive of financial results > Nonfinancial measures that are less predictive
- Objective measures > Subjective measures

Psychology-Based Hypotheses

- Outcome/results measures > Input/driver measures
- External measures > Internal measures
- Financial measures >/< Nonfinancial measures
- Objective/Quantitative measures > Subjective/Qualitative measures
- Measures based on multiple indicators > Single-item measures
- Measures with targets > Measures without targets

A > B: weight on measure A is greater than the weight on measure B.
A >/< B: weight on measure A has an ambiguous magnitude relative to the weight for measure B.

FIGURE 1
1992 Business Model for U.S./European Consumer Bank



replaced by the formula-based Performance Incentive Plan (PIP), which rewarded performance based on multiple, quantitative financial and nonfinancial measures. The PIP program's mechanics and evolution from 1993 to 1995 are summarized in Appendix A. Initially, branches were first required to receive satisfactory scores on any internal operational audits conducted during the quarter and to pass a customer satisfaction hurdle, as measured by a market research firm's survey of customer satisfaction with branch performance. Branches passing the customer satisfaction hurdle received quarterly bonuses for achieving improvement targets in any one of eight performance objectives related to growing the business (tier I and tier II household growth, consumer checking balance growth, business and professional checking balance growth, revenue growth, and relationship growth), resource management (expenses as a percent of revenue and footings as a percent of tier I and tier II households), and ''overall performance'' (quarterly margin growth).[3]

Minor changes occurred in 1994. In addition to passing the satisfaction hurdle and having satisfactory audit scores, branch managers were required to achieve targets in at least four of the eight performance objectives to receive a bonus. In 1995, the PIP objectives shifted further and included customer satisfaction (80 percent of customers rating overall satisfaction with GFS in the top two categories of the seven-point survey scale), growth (in tier I and tier II households, checking balances, liabilities and assets, and revenues), and resource management (growth in margins, and usage of automated tellers and other remote channels). To be eligible for bonuses, managers had to pass the customer satisfaction hurdle, have a satisfactory audit score, and meet their financial (revenue and margin) targets.

The growing complexity of the PIP bonus formula was reflected in the size of the document outlining each year's program: nine pages in 1993, 38 pages in 1994, and 78 pages in 1995. The primary cause of this complexity was upper management's frustration with a formula-based compensation system that allowed branch managers to game the system and earn bonuses without delivering financial results. To insure that branches were achieving financial targets, the 1995 PIP program added a financial hurdle that made it much more difficult for unprofitable branches to receive bonuses.

Balanced Scorecard Bonus Plan

Despite the increasing complexity of the PIP system, an internal evaluation concluded that results under the system had not put the organization on a trajectory to realize its strategic goals. In early 1995, the North American Banking Division (NABD) began replacing the Performance Incentive Plan with a ''balanced scorecard'' system focused on GFS's five corporate ''imperatives'' for success over time: achieving good financial results, delivering for customers, managing costs strategically, managing risk, and having the right people in the right jobs.[4] Scorecard implementation was accompanied by extensive training and widespread internal and external communication of the scorecard system's mechanics and objectives.

NABD's Western region replaced the PIP program with the balanced scorecard system in May 1995, with other NABD regions following in 1996. The research site provided us

[3] A household is a group that makes banking decisions as a family or business unit. Tier I households are customers with total combined balances in excess of $100,000 (including investment balances) and tier II households are customers with balances in excess of $10,000. Footings are defined as consumer and business/professional liabilities plus consumer and business/professional assets (excluding mortgages).

[4] Implementation of the scorecard was not limited to NABD. Each business unit in GFS was required to develop a scorecard that focused on the five corporate imperatives. More importantly, the company's training material and internal communications indicated that the firm expected business units to achieve superior performance on all five dimensions, using the slogan ''Five out of five, they all matter.''

with extensive data on the Western region's balanced scorecard performance measures and performance evaluations, branch manager compensation, financial performance, and related internal employee surveys. Internal documents and employee survey data were also provided for other GFS regions. We supplemented this material with extensive interviews of GFS personnel at all organizational levels, and with observations of balanced scorecard working groups, implementation teams, and bonus award meetings over the life of the scorecard plan.

Balanced Scorecard Measures

The performance measures in the Western region's balanced scorecard fell into six categories: financial, strategy implementation, customer, control, people, and standards. The first three categories were each measured using multiple *quantitative* indicators. Financial performance was evaluated based on revenues, expenses, and margins. Depending upon the quarter, between seven and 18 measures were used to evaluate strategy implementation. These measures primarily related to growth in the number of customers in different segments, customer attrition, and the level of assets under management (AUM) for each customer or customer segment.

Two measures evaluated customer-related performance: overall satisfaction with GFS and the branch quality index.[5] Control was measured by the results of periodic internal audits of operations and legal/regulatory compliance. The people and standards evaluations represented *qualitative* assessments by the branch managers' area director. Superiors were instructed to consider performance management, teamwork, training and development (both for the branch manager and other branch employees), and employee satisfaction when assessing people-related performance.[6] Standards criteria were leadership, business ethics and integrity, customer interaction and focus, community involvement, and contribution to the overall business.[7] In Table 2, we summarize the characteristics of the measures in the scorecard categories and relate these measures to the research hypotheses in Table 1.[8]

Bonus Determination Process

The Western region was organized into five geographic areas, each consisting of five to 20 branches.[9] Branch managers within these areas reported to an area director, who reported to the president of the Western banking operation. Unlike the formula-based PIP program, the balanced scorecard system required area directors to *subjectively* weight the various performance measures when evaluating branch managers' performance and determining their bonuses. This process is summarized in Figure 2. Area directors first compared

---

[5] See Appendix B for the questions used to compute the two customer satisfaction measures.

[6] Performance management was defined as a manager's ability to "achieve goals by coaching, motivating, empowering, hiring, supporting, promoting, recognizing, and challenging staff." Although employee satisfaction was considered in evaluating the people category, employee satisfaction surveys were not conducted on a regular basis, making the quarterly assessment of this measure qualitative. Moreover, there was no statistically significant correlation between the employee satisfaction scores from a 1996 survey and the subjective "people" scores given by area directors in the first and second quarters of 1996, indicating that quantitative employee satisfaction measures received little weight in evaluating managerial performance on this dimension.

[7] Formal targets were not provided for the control, people, and standards categories, but an audit rating of "3" or lower was "below par" performance in the control category.

[8] Definitions for all the variables used in our statistical tests are provided in Appendix B.

[9] A total of 95 branches were open at some point during the period covered by our study. The number of branches open in an individual quarter ranged from 76 to 89. Missing balanced scorecard data reduce the sample sizes in some of our quarterly tests. This is particularly true for 1997, when balanced scorecard data were only available for 45, 37, 21, and 36 branches in the four quarters, respectively. The explanation from GFS regarding the missing data was that the scorecards were lost and could not be retrieved by the company. We have no way to assess the degree to which these lost observations bias our results.

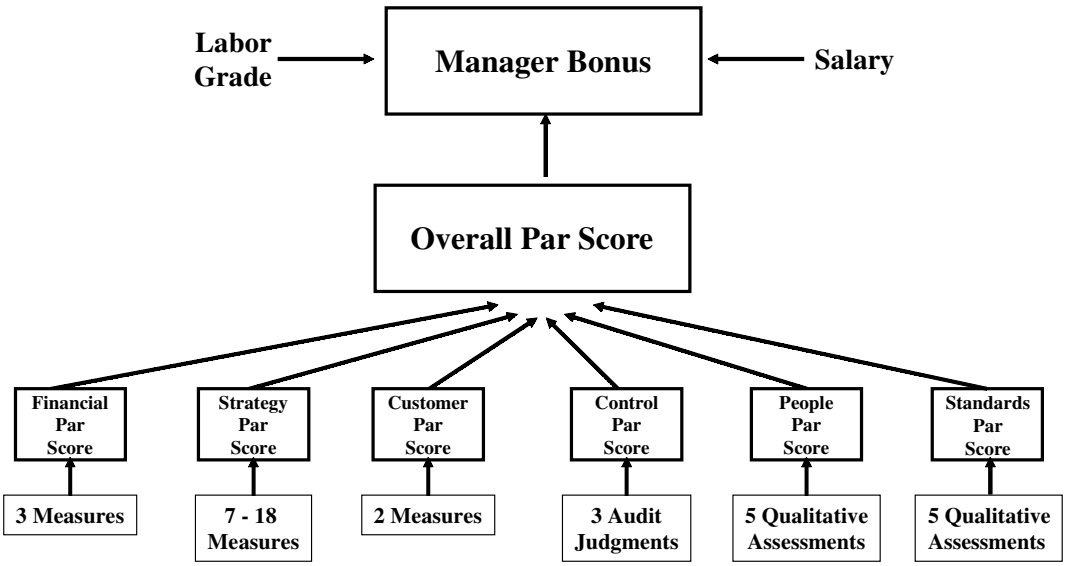
TABLE 2
Characteristics of GFS's Balanced Scorecard Measures and Link to Measurement Characteristics Used in the Research Hypotheses

	Balanced Scorecard Category					
Measurement Characteristic	**Financial**	**Strategy**	**Customer**	**Control**	**People**	**Standards**
Financial	x	Some				
Objective/Quantitative	x	x	x			
Subjective				x	x	x
Outcome measure	x	x				
Externally oriented	x	x	x			
Targets set	x	Some[a]	x	Audit hurdle		
Multiple measures	x	x	x			
Single overall assessment				x[b]	x[b]	x[b]

[a] Targets were set for all strategy implementation measures in the first four quarters under the balanced scorecard. In subsequent quarters, GFS set targets for only a subset of the strategy measures.

[b] Although area directors were instructed to consider multiple factors in assessing performance on these dimensions, only one overall assessment was given for each of these categories.

FIGURE 2
Balanced Scorecard Bonus Award Process



results to targets for each of the various financial, strategy implementation, and customer measures.[10] Branch managers then received a ''par rating'' for each of the measures within the financial, strategy, and customer categories, where ''below par'' reflected performance below expectations, ''at par'' represented expected performance, and ''above par'' reflected better-than-expected performance. Ratings for performance on the individual measures were then subjectively aggregated into ratings for the financial, strategy, and customer categories. The control, people, and standards categories each received a single *overall* par score (i.e., par scores were not given for the individual performance criteria within these three categories). The area director then subjectively combined the six scorecard categories into an overall performance rating of ''below par,'' ''at par,'' or ''above par.''

The quarterly bonuses recommended by the area directors were meant to reflect the branch manager's overall par score, labor grade, and current compensation. Unlike the PIP program, no formula was applied. Instead, bonuses were intended to achieve *total* market-based compensation levels (salary plus bonus) for a given labor grade and performance level. For example, assume that total compensation for branch managers in the highest of the three labor grades was targeted at *up to* $75,000 annually if performance was at par, *up to* $90,000 if performance was above par, and *up to* $105,000 or more if performance was exceptional. If a manager with an above par overall evaluation in this labor grade earned a salary of $80,000, then the *maximum* quarterly bonus was $2,500 ($10,000/4). However, if the manager's salary was $90,000 or more, no bonus was awarded despite the above par performance. This differed from the PIP formula, which determined a bonus percentage that was awarded regardless of the branch manager's salary (e.g., a branch

[10] Targets were only provided for a subset of strategy implementation measures after the fourth quarter under the plan.

manager with an annual salary of $80,000 and eligible for 15 percent bonus would receive a $3,000 bonus for that quarter, while a branch manager with an annual salary of $90,000 and eligible for a 15 percent bonus would receive a quarterly bonus of $3,375).

Area directors' bonus recommendations were taken to a meeting where the president of the Western region, the president's staff (the finance director, human resource director, compensation manager, and service quality director), and the five area directors discussed each recommendation. The discussion generally focused on the justification for the overall rating recommended for the branch manager, particularly when the overall evaluation of a manager was above par and the manager was eligible for a substantial bonus. Performance evaluations and bonuses could then be modified to promote consistency throughout the region.[11]

IV. RESULTS

Determinants of Scorecard Ratings

Our first set of tests addresses the weights placed on various types of performance measures when subjectivity is allowed in reward systems. We examine the *implicit* weights placed on the various performance measures each quarter by investigating the associations between the branches' performance on the scorecard measures, the par ratings given to branch managers in the six scorecard categories, and the size of their quarterly bonuses. These tests are similar to experimental studies using lens model methods to determine the relative weights placed on different types of information when evaluating performance, as well as archival compensation research that estimates the weights placed on accounting and stock market measures based on their statistical associations with compensation levels.

We first examine the regression weights placed on various *quantitative* measures.[12] Table 3 provides evidence on the associations between quarterly financial, customer, and strategy par ratings (where 1 = below par, 2 = at par, and 3 = above par) and performance versus targets for each of the quantitative performance measures in these categories.[13]

The financial par results in Panel A of Table 3 indicate that performance relative to revenue targets was factored into managers' evaluations during each of the quarters, with higher financial performance evaluations when revenues exceeded targets.[14] Expenses were negative in every quarter and statistically significant in ten of the 15 quarters ($p < 0.10$, two-tailed). In addition, the coefficients on revenues were substantially larger in absolute

[11] In contrast to the formula-based PIP system, which required no input by area directors, the scorecard process took area directors approximately six days per quarter to prepare the branch managers' scorecards, discuss the scorecards and bonus recommendations at the quarterly bonus meetings, and meet with the region's president. The extensive system of reviews under the scorecard plan is consistent with Prendergast and Topel's (1996) economic model, which indicates that extensive bureaucratic procedures are required to minimize the potential for favoritism and bias in subjective performance evaluations.

[12] For tests where the dependent variable is a par score (coded as 1, 2, or 3), we also conducted the statistical analysis using an ordered regression. These results were virtually identical to those reported in the tables.

[13] As noted earlier, GFS did not set targets for *all* of the strategy implementation measures after the first four quarters of the balanced scorecard. Consequently, our strategy par score analyses use absolute performance on measures without a goal for the later quarters and relative performance when goals are available. For measures with goals, we use the ratio of the actual outcome to the goal as the predictor variable.

[14] We delete observations from all regression tests when studentized residuals are greater in absolute value than 3 to mitigate the impact of outliers on our results.

TABLE 3
Regression Models Examining the Relation between Financial and Customer Par Scores and Their Underlying Quantitative Performance Measures

Par Score = f(Quantitative Performance Measures)

Panel A: Financial Par Scores

	Second Quarter 1995	Third Quarter 1995	Fourth Quarter 1995	First Quarter 1996	Second Quarter 1996	Third Quarter 1996	Fourth Quarter 1996	First Quarter 1997	Second Quarter 1997	Third Quarter 1997	Fourth Quarter 1997	First Quarter 1998	Second Quarter 1998	Third Quarter 1998	Fourth Quarter 1998
Revenue	8.01***	8.51***	4.73***	6.76***	5.30***	0.44*	7.00***	8.12***	6.44***	4.74***	4.03***	1.52***	1.26***	4.69***	3.01***
Expense	−1.26***	−2.14***	−4.32***	−2.56***	−1.61**	−0.77	−3.23***	−2.21**	−1.38	−3.17**	−0.80**	−0.36	−0.09	−0.91	−0.61**
Adj. R^2	0.64	0.57	0.49	0.56	0.49	0.05	0.62	0.57	0.37	0.49	0.30	0.18	0.14	0.41	0.31
n	73	74	83	89	89	81	80	45	37	21	36	74	81	66	58

Panel B: Customer Par Scores

	Second Quarter 1995	Third Quarter 1995	Fourth Quarter 1995	First Quarter 1996	Second Quarter 1996	Third Quarter 1996	Fourth Quarter 1996	First Quarter 1997	Second Quarter 1997	Third Quarter 1997	Fourth Quarter 1997	First Quarter 1998	Second Quarter 1998	Third Quarter 1998	Fourth Quarter 1998
GFS CSI	6.73***	7.83***	7.40***	7.64***	10.73***	9.02***	2.55***	7.60***	6.75***	7.52***	7.70***	7.78***	6.63***	6.35***	7.63***
Branch Quality Index	1.49***	0.41	−0.17	0.30	−0.84**	−0.41	1.15*	−0.17	−0.04	0.88*	0.19	−0.46	0.96**	0.17	−0.14
Adj. R^2	0.76	0.70	0.78	0.66	0.80	0.74	0.21	0.83	0.78	0.76	0.77	0.69	0.75	0.70	0.67
n	73	74	79	62	84	80	79	43	86	82	72	78	59	66	58

***, **, * Statistically significant at the 1%, 5%, and 10% levels, respectively (two-tailed).
See Appendix B for variable definitions.

value than those on expenses, suggesting that revenues received greater weight in performance evaluations. The mean (median) adjusted R^2 was 0.41 (0.49), implying that the quantitative performance measures for this category accounted for a little less than half of the managers' performance evaluation on this dimension.[15]

Nearly all the emphasis in customer-related evaluations was on overall satisfaction with GFS (Table 3, Panel B). The overall customer satisfaction measure was a significant determinant of customer par ratings in every quarter. The branch quality index, on the other hand, had a significant positive impact on customer par ratings in only four of the 15 quarters, and was significantly *negative* in the second quarter of 1996. Our interviews indicated that the primary reason for the heavy weight on overall satisfaction rather than the branch quality index was the fact that *all* senior managers and business units in GFS were evaluated using the common overall customer satisfaction measure. As a result, the Western region's president, who was also held responsible for this measure, drove this measure down to lower organizational levels in his division in order to provide an assessment of the region's progress toward the corporate goal of 80 percent overall satisfaction that was established for all GFS businesses. This result is consistent with Lipe and Salterio's (2000) experimental finding that evaluators place greater weight on common balanced scorecard measures than on unique measures, as well as contagion theories that suggest that lower-level managers evaluate subordinates using the same criteria used by upper-level managers to evaluate their performance (e.g., Hopwood 1974). In general, the customer par score models had substantially greater explanatory power than the financial models (mean adjusted $R^2 = 0.71$, median = 0.75)

Determinants of the strategy par scores are examined in Table 4. Given the large number of strategy performance measures (7 to 18 depending upon the quarter) and the small sample sizes in some of the quarters, we use stepwise regression models in these tests. The explanatory power of the strategy implementation measures varied widely from quarter to quarter, with adjusted R^2s ranging from 0.10 in the first quarter of 1996 to 0.72 in the third quarter of 1997 (mean = 0.32, median = 0.27). Moreover, no more than five of the measures were significant predictors of strategy par scores in any quarter ($p < 0.15$, two-tailed), and in many quarters only one or two of the strategy measures were significant (mean = 2.73 categories, median = 2).

The results in Tables 3 and 4 indicate that the subjective evaluations given to branch managers for the financial, customer, and strategic implementation categories were based only partially on the quantitative performance measures they were supposed to reflect. Adjusted R^2s ranged from 0.05 to 0.81, indicating that roughly 19 to 95 percent of a branch manager's par rating for a particular category was based on factors other than performance relative to the category's goals. Obviously, the explanatory power of these models is subject to the validity of the model specification linking the independent and dependent variables. Despite this caveat, the large unexplained variance in many of the models is consistent with analytical research indicating that the presence of subjectivity and discretion in reward systems allows superiors to base evaluations on factors other than the measures included in the compensation plan (Baker et al. 1994; Baiman and Rajan 1995; Prendergast and Topel 1996), even though this discretion was not intended by the scorecard plan's designers.

[15] We did not include margins (defined as revenues – expenses) in the model because of multicollinearity problems. When financial par scores were regressed on margins alone, the coefficients were positive and significant in each quarter. However, the adjusted R^2s for the models were substantially lower (mean = 0.22, median = 0.21). Low variance inflation factor (VIF) scores indicate no serious problems with multicollinearity in any of the other models reported in the paper.

TABLE 4
Stepwise Regressions Examining the Relation between Strategy Implementation Par Scores and Their Underlying Quantitative Performance Measures

Par Score = f(Quantitative Performance Measures)

Second Quarter 1995	Total AUM ($)^R	4.62***
Adj. R^2 = 0.41	Retail Households (#)^R	6.02***
n = 48	Household Attrition (#)^R	−1.74***
Third Quarter 1995	B&P Households (#)^R	3.11***
Adj. R^2 = 0.12	Total AUM ($)^R	2.18*
n = 57		
Fourth Quarter 1995	Retail Households (#)^R	6.92***
Adj. R^2 = 0.27	Total AUM ($)^R	0.93***
n = 73		
First Quarter 1996	B&P Households (#)^R	0.94***
Adj. R^2 = 0.10		
n = 75		
Second Quarter 1996	Retail Households (#)^R	0.42***
Adj. R^2 = 0.29	B&P CNR/Household ($)	0.00*
n = 58		
Third Quarter 1996	Total Cross Sales ($)	0.01***
Adj. R^2 = 0.34	B&P Households (#)^R	0.14*
n = 41		
Fourth Quarter 1996	New Premier Households (#)	0.04***
Adj. R^2 = 0.26	Retail Households (#)^R	0.27**
n = 46		
First Quarter 1997	Premier CNR/Household ($)^R	0.56**
Adj. R^2 = 0.41	Lost Retail Households (#)	−0.01***
n = 43	New Retail Households (#)	0.01**
Second Quarter 1997	B&P CNR/Household ($)^R	1.69***
Adj. R^2 = 0.63	New Premier Households (#)	0.13***
n = 36	Lost B&P Households (#)	−0.04***
	New B&P Households (#)	0.02**
	Retail Households (#)^R	0.94*
Third Quarter 1997	Automated Transactions (% of total)	0.03***
Adj. R^2 = 0.72	Premier CNR/Household (#)^R	0.81***
n = 21	Lost B&P Households (#)	−0.04***
	New Retail Households (#)	0.01**
	B&P CNR/Household ($)^R	0.79*
Fourth Quarter 1997	New Premier Households (#)	0.05***
Adj. R^2 = 0.24		
n = 25		
First Quarter 1998	Retail Households (#)^R	1.64***
Adj. R^2 = 0.23	New Retail Households (#)	0.03***
n = 70	Lost Retail Households (#)	−0.00*
Second Quarter 1998	New B&P Households (#)	0.06***
Adj. R^2 = 0.32	Retail Households (#)^R	1.21***
n = 52	Lost Retail Households (#)	−0.00#
	New Retail Households (#)	0.01*
	Premier Households (#)^R	0.52*
Third Quarter 1998	Premier CNR/Household ($)^R	0.71***
Adj. R^2 = 0.23	Premier Households (#)^R	0.49**
n = 59		
Fourth Quarter 1998	Premier CNR/Household ($)^R	0.35***
Adj. R^2 = 0.22	B&P CNR/Household ($)^R	0.91***
n = 53	Retail Households (#)^R	1.18*

***, **, *, # Statistically significant at the 1%, 5%, 10%, and 15% levels, respectively (two-tailed).
^{R} Performance measure is relative to goal.
See Appendix B for variable definitions.

Association between Ratings in Individual Scorecard Categories and Overall Ratings

Our second set of tests investigates the association between the qualitative ratings given in each of the six scorecard categories and the branch manager's overall performance rating. Since the par scores were categorical variables (below par, par, or above par), we estimate general linear models with indicator variables for the three par score categories. We evaluate the proportion of variance in the dependent variable attributable to each par score using partial eta squareds (η^2).[16]

The results in Table 5 indicate that the financial and customer par scores were the most consistent determinants of overall performance ratings. Partial eta squareds for the financial par scores were significant in all but one of the quarters ($p < 0.10$, two-tailed F-test), while customer par scores were significant in 12 of the quarters. In contrast, the qualitative standards, people, and control par scores were only significant determinants of overall par scores in five, four, and two quarters, respectively. Surprisingly, the strategy par score, which was intended to be based on quantitative performance indicators, was only significant in six of the 15 quarters, a result similar to the qualitative standards assessment. This evidence is inconsistent with our hypothesis that quantitative performance measures or assessments based on multiple indicators receive greater weight in performance evaluations, but is consistent with studies finding lower weight on measures when performance targets (which were not set for all of the strategy measures after the first four quarters under the plan) are not provided (e.g., Heneman 1986; Bommer et al. 1995).

To examine whether the magnitudes of the partial eta squareds were significantly different among the performance categories, we compute nonparametric Wilcoxon signed rank tests between each pair of categories (not reported in the tables). Median eta squareds for the financial par scores were significantly larger than those for any of the other performance categories ($p < 0.01$, two-tailed). Median eta squareds for the customer par scores were also significantly larger than medians for strategy implementation and the three qualitative categories. Strategy implementation eta squareds, on the other hand, were only marginally larger than those for the control and standards categories ($p < 0.15$, two-tailed), and were not significantly different than the people eta squareds.

These analyses indicate that financial par scores were used more frequently and received greater weight than nonfinancial par scores. As hypothesized, both the financial and customer par scores, which were based on externally oriented, quantitative results measures, received greater emphasis than the more qualitative, internally oriented customer, people, and control par scores. However, the limited emphasis placed on strategy par scores is inconsistent with our hypotheses that quantitative measures and subjective evaluations based on multiple criteria receive greater weighting. Although strategy par scores were meant to reflect their underlying quantitative measures (which the research site's business model assumed to be key externally oriented outcome measures), we find that they were significant determinants of overall par scores in only 40 percent of the quarters, and generally received weights similar to the single-item qualitative performance evaluations.

Finally, even though nonfinancial performance was incorporated into the area director's overall assessment of a branch manager, the significantly larger eta squareds for the financial par scores indicate that the balanced scorecard system was primarily driven by financial

[16] The partial eta squared is the proportion of the effect plus error variance that is attributable to an effect, and is calculated as: $\eta_p^2 = SS_{effect}/(SS_{effect} + SS_{error})$. The sums of the partial eta squared values are not additive, and do not sum to the amount of dependent variable variance accounted for by the independent variables. Furthermore, it is possible for the sums of the partial eta squareds to be greater than 1.

TABLE 5
Partial Eta Squareds from GLM Models Examining the Relation between Overall Par Scores and Individual Par Scores

Overall Par Score = f(Individual Par Scores)

	Second Quarter 1995	Third Quarter 1995	Fourth Quarter 1995	First Quarter 1996	Second Quarter 1996	Third Quarter 1996	Fourth Quarter 1996	First Quarter 1997	Second Quarter 1997	Third Quarter 1997	Fourth Quarter 1997	First Quarter 1998	Second Quarter 1998	Third Quarter 1998	Fourth Quarter 1998
Financial	0.42***	0.28***	0.56***	0.63***	0.53***	0.37***	0.21***	0.44***	0.66***	0.35	0.33*	0.43***	0.37***	0.30***	0.41***
Strategy	0.02	0.14***	0.13***	0.08	0.12**	0.01	0.02	0.04	0.20	0.01	0.05	0.28***	0.29***	0.26**	0.12
Customer	0.11**	0.10**	0.06	0.24***	0.50***	0.24***	0.16***	0.01	0.34**	0.27	0.36*	0.30***	0.33***	0.38***	0.24*
Control	0.02	0.06**	0.02	0.32***	0.09*	0.01	0.01	0.18*	0.00	0.05	0.09	0.12*	0.14*	0.09	0.00
People	0.00	0.14***	0.04	0.16**	0.06	0.08	0.27***	0.07	0.01	0.13	0.03	0.30***	0.01	0.24**	0.09
Standards	0.08*	0.06	0.08**	0.02	0.01	0.08*	0.01	0.27**	0.02	0.06	0.00	0.05	0.09*	0.02	0.02
Model Adj. R^2	0.50	0.55	0.75	0.87	0.79	0.60	0.53	0.74	0.72	0.38	0.65	0.67	0.75	0.73	0.71
n	73	72	80	61	70	53	69	42	26	20	25	56	41	42	31

***, **, * Statistically significant at the 1%, 5%, and 10% levels, respectively (two-tailed).
See Appendix B for variable definitions.

considerations. This evidence suggests that the balanced scorecard plan only partially addressed criticisms that GFS's earlier formula-based compensation program provided incentives for branch managers to focus their efforts on a single or limited set of performance dimensions.

Associations between Scorecard Ratings and Quarterly Bonuses

We examine the extent to which the performance ratings in the six scorecard categories were weighted in bonus computations in Table 6. Because the maximum bonus award under the scorecard was intended to be a function of the branch manager's labor grade and current salary, we include these factors as control variables in the analyses (where 1 = the lowest labor grade and 3 = the highest).

The partial eta squareds in Panel A of Table 6 investigate the relation between bonuses and branch managers' overall performance ratings. Results for labor grade and salary are not presented to simplify presentation.[17] The overall par score is a significant predictor of bonuses in every quarter ($p < 0.10$. two-tailed), with the models' explanatory power ranging from 0.19 to 0.70 (mean = 0.48, median = 0.45).

The influence of the individual scorecard categories on bonus awards is examined in Panel B of Table 6. Financial performance had the most frequent direct association with bonus awards. The financial par score was significant in nine of the 15 quarters, with a mean (median) partial eta squared of 0.27 (0.29). The other par scores were only significant in four or fewer quarters. This evidence suggests that any effects of nonfinancial measures on bonus awards were primarily through their influence on overall par scores, rather than through their direct influence on bonuses.

Taken together, the results in Tables 5 and 6 suggest that bonus computations under the balanced scorecard system were focused much more on financial outcomes than were bonus computations under the formula-based PIP system. Under the PIP system, the performance measures in the compensation plan could not be ignored as long as the manager met the appropriate hurdles and targets. In contrast, the results in Tables 3 and 4 indicate that a number of the performance measures in the scorecard had little or no influence on performance evaluations and bonus awards. Moreover, unlike the strict formula used to compute bonuses under the PIP system, the balanced scorecard performance measures explained only about 50 percent of the observed bonuses. The large unexplained variance in bonus amounts is consistent with the greater discretion in the balanced scorecard system than in the PIP system, which did not allow deviations from the bonus formula.

Associations between Nonfinancial Measures and Financial Performance

One potential explanation for the limited weights placed on some of the nonfinancial measures is that these measures contained little incremental information on managerial actions. Many economics-based agency models indicate that performance measures should only be included in compensation plans when they provide incremental information on the actions desired by the principal. A key assumption of GFS's scorecard system was that the nonfinancial scorecard measures were *leading* indicators of future financial results. If some of the nonfinancial balanced scorecard measures actually had no ability to predict future financial performance, their omission in performance evaluations could be due to area directors ignoring or abandoning uninformative performance indicators.

[17] As expected, bonuses were smaller at lower labor grades and at higher salary levels (after controlling for labor grade) due to the maximum targeted total compensation levels established for each labor grade.

TABLE 6
Partial Eta Squareds from GLM Models Examining the Relations between Bonus Awards and Par Scores

Panel A: Bonus Awards = *f*(Overall Par Score, Salary, Labor Grade)

	Second Quarter 1995	Third Quarter 1995	Fourth Quarter 1995	First Quarter 1996	Second Quarter 1996	Third Quarter 1996	Fourth Quarter 1996	First Quarter 1997	Second Quarter 1997	Third Quarter 1997	Fourth Quarter 1997	First Quarter 1998	Second Quarter 1998	Third Quarter 1998	Fourth Quarter 1998
Overall Par Score	0.24***	0.69***	0.33***	0.57***	0.44***	0.24***	0.34***	0.34***	0.62***	0.12	0.42***	0.29***	0.62***	0.15**	0.25***
Model Adj. R^2	0.38	0.70	0.61	0.60	0.55	0.38	0.33	0.31	0.67	0.45	0.45	0.32	0.62	0.19	0.57
n	56	64	63	61	75	69	73	41	31	18	35	70	57	62	56

Panel B: Bonus Awards = *f*(Individual Par Scores, Salary, Labor Grade)

	Second Quarter 1995	Third Quarter 1995	Fourth Quarter 1995	First Quarter 1996	Second Quarter 1996	Third Quarter 1996	Fourth Quarter 1996	First Quarter 1997	Second Quarter 1997	Third Quarter 1997	Fourth Quarter 1997	First Quarter 1998	Second Quarter 1998	Third Quarter 1998	Fourth Quarter 1998
Financial	0.30***	0.20**	0.47***	0.39***	0.32***	0.45***	0.23***	0.17	0.51**	0.29	0.21	0.09	0.10	0.02	0.36**
Strategy	0.10	0.23*	0.08	0.04	0.12*	0.06	0.05	0.09	0.06	0.41	0.24	0.06	0.21**	0.12	0.03
Customer	0.08	0.09	0.06	0.03	0.02	0.00	0.13**	0.01	0.15	0.12	0.46**	0.07	0.17*	0.02	0.14
Control	0.00	0.01	0.01	0.02	0.03	0.08	0.09	0.00	0.15	0.44	0.15	0.08	0.07	0.08	0.09
People	0.01	0.06	0.00	0.02	0.01	0.12*	0.03	0.08	0.02	0.02	0.08	0.01	0.08	0.01	0.09
Standards	0.02	0.02	0.10**	0.00	0.03	0.00	0.04	0.15	0.00	0.05	0.09	0.00	0.00	0.05	0.07
Model Adj. R^2	0.41	0.37	0.62	0.56	0.63	0.59	0.40	0.39	0.52	0.54	0.58	0.21	0.54	0.06	0.62
n	49	55	61	49	64	48	63	40	23	17	25	54	41	42	31

***, **, * Statistically significant at the 1%, 5%, and 10% levels, respectively (two-tailed).
Salary and labor grade results are not presented to simply presentation.
See Appendix B for variable definitions.

We provide some evidence on the informativeness of the nonfinancial measures by examining whether changes in these measures were leading indicators of the bank's two strategic objectives (customer growth and financial performance). Consistent with the business model in Figure 1, we regress percentage changes in financial performance or customer growth on percentage changes in the quantitative customer measures and the qualitative people, control, and standards par scores. We also include the customer growth measures as predictor variables in the financial performance models to examine whether changes in customer growth were associated with current or future changes in financial results. Indicator variables for time period are included as control variables to account for potential time-specific effects on performance.[18]

We examine these relations using three time lags. The first two tests examine short-term effects using contemporaneous changes and one quarter lags (e.g., percentage changes in the dependent variables between quarters 2 and 3 are regressed on percentage changes in the predictor variables between quarter 1 and 2). Although these lags are relatively short, the frequent repurchase cycle and relatively low customer switching costs in retail banking can lead to reasonably short lags between managerial actions and observed economic performance. Longer-term effects are examined using rolling four quarter lags. For example, percentage changes in financial performance and customer growth are computed using the current quarter in a given year and the same quarter one year earlier (e.g., quarter 1 results for 1998 divided by quarter 1 results for 1997). These changes are regressed on rolling percentage changes in the predictor variables during the previous four quarters (e.g., quarter 1 results for 1997 are divided by quarter 1 results for 1996). By computing percentage changes relative to the same quarter in the prior year, we control for seasonality.

The results are presented in Table 7. We find little evidence consistent with the hypothesis that the weights placed on nonfinancial measures are a function of their ability to predict future changes in performance (our proxy for informativeness). Although overall satisfaction with GFS was the primary determinant of customer par scores, this measure had no contemporaneous relation with customer growth or financial performance, and was *inversely* related to some of the future results. In particular, changes in GFS satisfaction were negatively related to changes in Retail and Premier customers and positively related to expense changes in the following quarter, and were negatively related to changes in Retail customers and margins in the subsequent four quarters. The associated reductions in Retail customers also appear to have had a negative *indirect* effect on margins due to the positive contemporaneous and four quarter lagged relations between Retail customers and margins and the positive relation with expense changes in the following quarter.

In contrast, the branch quality index, which received relatively little weight in the determination of customer par scores, had a significant positive association with four quarter lagged changes in margins but not with short-term margin changes, suggesting that improvements in branch satisfaction only yielded improvements in financial performance with some delay. The branch quality index was also positively associated with contemporaneous changes in retail customers, and negatively associated with expense changes in the subsequent quarter. Percentage changes in the number of retail customers, in turn, had positive relations with current revenues and margins and with four quarter lagged margins. Taken together, these results are consistent with branch quality having positive direct effects on

[18] We also estimated these models with indicator variables for each of the five distinct regions in an attempt to control for potential regional differences. These results were virtually identical to those reported in Table 7.

TABLE 7
Regression Models Examining the Relation between Percentage Changes in Customer Growth or Financial Performance and Percentage Changes in Nonfinancial Performance Measures

%ΔCustomer Growth or *%ΔFinancial Performance = f(%ΔNonfinancial Performance Measures, Time Period Indicators)*

Panel A: Contemporaneous Relations

	%ΔRetail Customers	*%ΔB&P Customers*	*%ΔPremier Customers*	*%ΔRevenue*	*%ΔExpenses*	*%ΔMargin*
Intercept	0.933***	1.123***	1.616***	0.833***	1.092***	0.443**
%ΔGFS CSI	−0.015	−0.026	0.099	0.025	−0.076	0.121
%ΔBranch Quality Index	0.086**	0.004	−0.026	0.006	−0.011	−0.058
%ΔControl Par Score	0.019	0.010	0.021	−0.019#	0.020	−0.022
%ΔPeople Par Score	−0.012	−0.038	−0.100*	0.043**	−0.045**	0.148***
%ΔStandards Par Score	−0.009	0.035	0.045	−0.011	0.000	−0.074
%ΔRetail Customers				0.113**	0.003	0.413***
%ΔB&P Customers				0.004	0.008	0.015
%ΔPremier Customers				0.006	0.012	0.009
Adj. R^2	0.058	0.006	0.235	0.096	0.309	0.182
F-statistic	2.583***	1.145	8.154***	3.129***	9.998***	5.419***
n	461	460	420	421	424	417

Panel B: One Quarter Lag

	%ΔRetail Customers	*%ΔB&P Customers*	*%ΔPremier Customers*	*%ΔRevenue*	*%ΔExpenses*	*%ΔMargin*
Intercept	1.189***	1.219***	1.393***	0.995***	1.102***	1.007***
%ΔGFS CSI	−0.093#	−0.068	−0.275*	−0.018	0.118*	−0.135
%ΔBranch Quality Index	−0.052	−0.053	0.039	0.031	−0.086*	0.111
%ΔControl Par Score	−0.015	0.008	−0.013	0.017	−0.017	−0.006
%ΔPeople Par Score	0.006	0.013	−0.011	−0.001	0.054***	−0.079*
%ΔStandards Par Score	0.015	−0.025	0.059	0.016	0.011	0.087

(*continued on next page*)

TABLE 7 (continued)

Panel B: One Quarter Lag (continued)

	%ΔRetail Customers	*%ΔB&P Customers*	*%ΔPremier Customers*	*%ΔRevenue*	*%ΔExpenses*	*%ΔMargin*
%ΔRetail Customers				0.003	0.011*	−0.006
%ΔB&P Customers				−0.003	−0.004	−0.006
%ΔPremier Customers				−0.007	0.027***	−0.025#
Adj. R^2	0.054	0.019	0.241	0.071	0.454	0.110
F-statistic	2.387***	1.462*	8.134***	2.420***	16.516***	2.997***
n	412	412	382	372	374	372

Panel C: Four Quarter Lag

	%ΔRetail Customers	*%ΔB&P Customers*	*%ΔPremier Customers*	*%ΔRevenue*	*%ΔExpenses*	*%ΔMargin*
Intercept	1.148***	1.283**	0.458	1.163*	1.317***	0.879***
%ΔGFS CSI	−0.136*	−0.169	0.692	−0.151	0.036	−0.438#
%ΔBranch Quality Index	−0.007	0.092	0.128	0.071	−0.107	0.411**
%ΔControl Par Score	−0.009	−0.046	0.096	−0.021	0.024	−0.076
%ΔPeople Par Score	−0.009	−0.002	0.395	−0.031	0.061**	−0.132**
%ΔStandards Par Score	0.002	0.024	0.191	−0.057	−0.019	−0.091
%ΔRetail Customers				0.107	−0.029	0.452**
%ΔB&P Customers				0.115**	−0.024	0.117
%ΔPremier Customers				0.001	0.004	−0.002
Adj. R^2	0.061	0.045	0.054	0.088	0.345	0.089
F-statistic	2.217***	1.892**	2.079***	2.331***	8.211***	2.333**
n	207	211	209	194	193	192

***, **, *, # Statistically significant at the 1%, 5%, 10%, and 15% levels, respectively (two-tailed).
Indicator variables for time period are included in the models but not reported to simplify presentation.
See Appendix B for variable definitions.

future margins, as well as positive indirect effects on current and future financial performance through growth in retail customers.

The other two customer-growth measures show little positive association with financial performance. B&P customer growth was associated with *lower* margins in the following month. Similarly, the percentage change in the number of Premier households, which was a significant determinant of strategy par scores in many quarters, was also *negatively* associated with margins in the following month due to the higher expenses to acquire and maintain these customers. Neither Premier nor B&P customer growth was associated with changes in financial performance over the next four quarters, despite the fact that measures related to these customer groups were significant determinants of strategy par scores in 12 of the 15 quarters.

Finally, changes in the qualitative control and standards par scores exhibit few significant associations with current or future performance, even though these measures had some influence on overall par scores. People par score changes had the expected positive associations with contemporaneous financial performance (higher revenue, lower expenses, and higher margins), but were *negatively* associated with subsequent (one-quarter and four-quarters ahead) margins due to higher expenses.

Although the results in Table 7 are subject to a variety of econometric concerns such as correct model specification, they are generally inconsistent with informativeness theories on the choice of performance measures. In the case of the branch quality index, the limited use of this measure appears to reflect the omission of value-relevant information from performance evaluations rather than the exclusion of a measure that provided no information on managerial performance. In other cases, measures that did not exhibit the statistical associations with financial performance predicted by GFS's business model received significant weight in performance evaluations. This is particularly true of the overall GFS customer satisfaction measure, which was *negatively* associated with the bank's strategic objectives. Together with the earlier results, this evidence suggests that psychology-based explanations may be equally or more relevant than economics-based explanations in explaining measurement practices at GFS.

V. QUALITATIVE ANALYSES

A summary of the preceding hypothesis tests is presented in Table 8. These statistical tests indicate that the subjectivity in the balanced scorecard plan allowed area directors to incorporate factors other than the scorecard measures in performance evaluations, to change evaluation criteria from quarter to quarter, to ignore measures that were predictive of future financial performance, and to weight measures that were not predictive of desired results. However, these analyses shed little light on whether the resulting scorecard plan achieved its objectives.

Some evidence on the outcomes of the balanced scorecard program can be obtained from two internal branch manager surveys conducted by GFS's human resource department.[19] Table 9 examines the balanced scorecard's influence on Western branch managers' perceived understanding of strategic goals and their attitudes toward performance evaluation criteria using data from internal employee surveys conducted in October 1994 (under the

[19] Sample sizes were 77 for the 1994 survey and 83 for the 1996 survey. Response rates were 97 percent and 93 percent, respectively.

TABLE 8
Summary of Hypothesis Test Results

Hypothesis regarding Relative Weights	Results
Economics-Based Hypotheses	
Nonfinancial measures that are more predictive of financial results > Nonfinancial measures that are less predictive	Not consistent (Table 7)
Objective measures > Subjective measures	Consistent for financial and customer measures; not consistent for strategy measures (Tables 5 and 6)
Psychology-Based Hypotheses	
Outcome/Results measures > Input/Driver measures	Consistent for financial measures; not consistent for strategy measures (Tables 5 and 6)
External measures > Internal measures	Consistent for financial and customer measures; not consistent for strategy measures (Tables 5 and 6)
Financial measures >/< Nonfinancial measures	Greater weight on financial measures (Tables 5 and 6)
Objective/Quantitative measures > Subjective/Qualitative measures	Consistent for financial and customer measures; not consistent for strategy measures (Tables 5 and 6)
Measures based on multiple indicators > Single-item measures	Consistent for financial and customer measures; not consistent for strategy measures (Tables 5 and 6)
Measures with targets > Measures without targets	Consistent (Tables 5 and 6)

PIP program) and February 1996 (under the balanced scorecard system). The mean responses in 1994 and 1996 reveal few statistical differences in perceptions under the formula-based PIP program and the more subjective, but broader, balanced scorecard. The scorecard's implementation brought little change in branch managers' stated understanding of strategic goals or their connection to the managers' actions. Under both systems, branch managers, on average, claimed that they understood GFS's business goals, the goals of their work group, the connection between their jobs and the business objectives, and the basis on which performance was judged. In contrast, the managers generally agreed that the GFS strategy had become clearer to them between 1995 and 1996, suggesting that the scorecard may have helped to communicate the company's strategic goals. However, branch managers felt less comfortable with the adequacy of the information provided to them about progress toward the multiple business goals. The perceived importance of customers and employee development in performance evaluation and compensation decisions, as well as agreement with the statement ''compensation decisions are consistent with performance,'' were not significantly different under the two systems.

The responses in Table 9 provide little evidence that the change from the PIP system to the balanced scorecard had an impact (either positive or negative) on managerial perceptions of business strategies, goals and priorities, performance evaluation, and compensation bases, or the adequacy of measures for decision making. However, the survey was conducted only three full quarters after the start of the scorecard system and was not focused on assessing the scorecard's impact. A subsequent survey covering a broader sample of

TABLE 9
Mean Employee Survey Responses by Western Region Branch Managers under the Performance Incentive Plan and Balanced Scorecard

	Performance Incentive Plan 1994 (n = 77)	Balanced Scorecard 1996 (n = 83)
A. Understanding of Strategy and Business Objectives		
I understand the business goals of GFS[a]	1.83	1.75
During the past year, the GFS strategy has become clearer to me[a]	NA	2.02
Senior management has communicated a clear plan for meeting our business goals[a]	2.13	2.32
I see the connection between the business objectives and my job[a]	1.93	1.71
B. Goal Setting		
I understand the goals of my work group[a]	NA	1.77
I know the basis on which my performance will be judged[a]	1.84	1.87
I get adequate information about progress against business goals[a]	2.00	2.87*
C. Performance Evaluation and Compensation		
Service to the customer is an important part of the way my performance is measured[a]	1.75	1.79
I am recognized for the service I provide to customers[a]	2.18	2.48
Managers are rewarded for developing their employees[a]	2.72	2.90
Decisions about my compensation have been consistent with my performance[a]	2.82	2.85

* Significantly different from the mean 1994 survey response at the 10% level or better (two-tailed).
[a] 1 = strongly agree, 2 = agree, 3 = neither agree nor disagree, 4 = disagree, 5 = strongly disagree.

North American managers was conducted in 1997 to specifically address the implementation of the scorecard.[20] Respondents were asked to use a three-point scale (agree, neutral, or disagree) to answer questions on their understanding of the scorecard process, implementation and goal setting, performance evaluation and bonus awards, and overall assessment of the scorecard process.

The results from the second survey are presented in Table 10. Although the majority of managers stated that they understood the scorecard process and their scorecard goals, only 32 percent of respondents were satisfied with the scorecard process, while 45 percent were dissatisfied. Three statements that managers particularly disagreed with were the scorecard process fairly assessing job performance (48 percent disagreeing versus 31 percent agreeing), the scorecard goals covering all important parts of the job (40 percent disagreeing

[20] Survey responses were received from 572 managers, representing a 74 percent response rate.

TABLE 10
1997 Survey Responses by 572 North American Managers on the Balanced Scorecard Process

	Agree	Neutral	Disagree
A. Understanding of the Scorecard Process			
I have a good understanding of the scorecard process	61%	18%	21%
My manager has a good understanding of the scorecard process	71	20	9
With scorecard, it is easy to see the connection between individual and branch performance	41	28	31
B. Implementation and Goal Setting in the Scorecard Process			
The scorecard process is too cumbersome to use effectively	25	31	44
The scorecard goal-setting process is too time consuming to do an adequate job	24	33	43
I am informed about the scorecard goals set for me	68	13	18
I know how my scorecard goals are determined	60	16	24
My scorecard goals can be realistically met, if I work effectively	40	28	33
My scorecard goals cover all the important parts of my job	39	21	40
C. Performance Evaluations and Bonus Awards Using the Scorecard			
The scorecard process fairly assesses my job performance	31	21	48
Bonuses given to people in my unit accurately reflect differences in their performance	30	35	35
When it comes to scorecard bonuses, I have no idea who gets what and why	55	20	25
For employees at my level, bonuses are higher under scorecard than they were under PIP	14	48	39
D. Overall Assessment of the Scorecard Process			
My Area Director believes in the usefulness of the scorecard process	58	38	4
Overall I am satisfied with the scorecard process	32	23	45

versus 39 percent agreeing), and bonuses accurately reflecting differences in their performance (35 percent disagreeing versus 30 percent agreeing).

These attitudes were also reflected in responses given in the open-ended, write-in portion of the survey. The most common issues addressed in this section related to the scorecard being incomplete or overly focused on financial performance (21 percent of the write-in responses). Representative comments included the following:

> I feel we are rated solely on revenues due to the pressure to increase our stock price. No 80 percent employee satisfaction first and 80 percent customer satisfaction second. Revenue is always first and only.

> As it stands now, it [expletive deleted]. Scorecard should be more heavily weighted as to service, relationship development, and contributions to the team. ''Product pushers'' are not the only way to win customers.

> Still too much emphasis on ''numbers'' and not on a ''balanced'' scorecard.

> Our job function entails so much more than the scorecard covers. The scorecard gives a very, very limited picture.

The second most frequent issue related to the inability of managers to understand how their bonuses were determined (14 percent of responses). For example:

> This is a black box process, no one knows anything.
>
> It's too subjective and not objective. I'd prefer an objective rating where everyone concerned knows what to expect when certain levels of performance are achieved.
>
> I would have liked to have known what my evaluation was going to be based on before the quarter began.
>
> I don't understand how it works! I don't see anyone motivated by it. I see the opposite.

Other frequent complaints centered on perceived favoritism, bias, or excessive discretion in scorecard bonus awards (12 percent of responses). Typical comments included the following:

> Eliminate the scorecards! Promotions and bonuses are still given to those who kiss-up to their supervisors, with little regard to performance, educational background, and experience.
>
> I hate the new process. Favoritism comes too much into play.
>
> Some people will get more money, regardless if they do a good job or not, just because they are someone's favorite. Discretion at the AD [area director] level must be eliminated to ensure fairness.

These qualitative analyses indicate that the balanced scorecard at GFS did not achieve its objectives, with GFS experiencing many of the problems identified in behavioral and economic models on the use of subjectivity in reward systems (Prendergast and Topel 1993, 1996). In response, senior management commissioned a team of business unit managers and human resource and compensation staff to review the balanced scorecard plan. The team recommended that the balanced scorecard concept be retained, but that incentive pay be divided into two components: (1) a ''structured'' or formula-based component, paid quarterly, based on quantitative measures including branch revenues and margins, customer satisfaction indicators, and household growth; and (2) a ''discretionary'' component, paid annually, based on more qualitative performance objectives such as people leadership, community involvement, projects, and teamwork. By separating bonuses into two distinct segments, the team attempted to retain the benefits of subjectivity while making the link between performance on the quantitative measures and bonus payouts more objective and transparent. Despite this recommendation, the company decided to return to a formula-based bonus plan that rewarded branch managers solely on the basis of branch revenues.

VI. CONCLUSIONS

In this study we examine how different types of performance measures (e.g., financial versus nonfinancial, quantitative versus qualitative, drivers versus results) are weighted in subjective bonus computations. We find that the use of subjectivity in weighting the measures in a balanced scorecard bonus plan allowed supervisors to ignore many performance measures, with financial performance became the primary determinant of bonuses. In addition, the subjectivity in the balanced scorecard plan allowed area directors to incorporate factors other than the scorecard measures in performance evaluations, to change evaluation criteria from quarter to quarter, to ignore measures that were predictive of future financial performance, and to weight measures that were not predictive of desired results. These

outcomes led many branch managers to complain about favoritism in bonus awards and uncertainty in the criteria being used to determine rewards, and caused corporate executive and human resource managers to question the scorecard's use for compensation purposes. Ultimately, the balanced scorecard was eliminated in the retail banks, along with many supporting measures such as customer satisfaction and branch quality.

Although it is difficult to generalize results from a single research site, our study provides a number of implications for future research. First, the evidence suggests that psychology-based explanations may be equally or more relevant than economics-based explanations in understanding measurement practices in some settings. In contrast to economic theories, we find little evidence that the weights placed on nonfinancial measures had any association with their ability to predict financial performance (our proxy for informativeness). Instead, we find that most of the weight was placed on quantitative, outcome-oriented financial measures that were used in earlier bonus plans, results that are consistent with psychology-based predictions. This evidence complements prior empirical research indicating that factors other than informativeness influence performance measure choices (e.g., Merchant 1989; Ittner and Larcker 2002).

Second, the study highlights the evolutionary nature of compensation plans. This is an important issue that is missing in pure cross-sectional studies, thereby limiting our understanding of how firms redesign their reward systems in response to unanticipated consequences. Third, our detailed examination of GFS's scorecard plan provides evidence on the multifaceted and complex nature of subjectivity. Prior empirical and theoretical studies have typically examined one element of subjectivity at a time (e.g., flexibility in weighting quantitative performance measures when computing a manager's bonus, the use of qualitative performance evaluations, or the discretion to adjust bonus awards based on factors other than the measures specified in the bonus contract). In contrast, our evidence indicates that all of these elements of subjectivity can occur simultaneously. Further analysis of the use and consequences of subjectivity will need to consider the potential simultaneous presence of these different elements. Finally, our quantitative and qualitative analyses indicate that implementation issues may be far more important to the success or failure of balanced scorecard systems than the scorecard's technical attributes (e.g., the number and types of measures, their classification into categories, or the presence of a causal business model). Consequently, future research on scorecard adoption and performance consequences must move beyond the measurement of these attributes to encompass the entire implementation process.

APPENDIX A
Evolution of the Formula-Based Performance Incentive Plan (PIP) System

Year	Hurdles	Performance Objectives	Bonus for Meeting Performance Targets	Additional Bonus for Exceeding Performance Targets
1993	Satisfaction with primary branch—top 75% of the Western region	Margin growth	3%	—
		Tier I and II household growth	2%	—
		Consumer checking balance growth	2%	—
		B&P checking balance growth	2%	—
		Revenue growth	2%	—
		Liability relationship growth	2%	—
		Expense control	1%	—
		Expenses/revenues	1%	—
1994	Satisfaction with primary branch—statistically at or above the region mean	Margin growth	3%	Two steps, to 1.5%
		Tier I and II household growth	1.5%	Two steps, to 2.5%
	Operations control—audit score of "4" or "5"	Consumer checking balance growth	1.5%	Two steps, to 2.5%
		B&P checking balance growth	1.5%	Two steps, to 2.5%
		Revenue growth	3%	Two steps, to 4.5%
		Liability relationship growth	1.5%	Two steps, to 2.5%
		Expenses/revenues	0.5%	Two steps, to 1%
		Footings/tier I and II households	0.5%	Two steps, to 1%
1995	Branch quality index—at or above the region mean	Overall GFS satisfaction ≥ 80%	5%	—
		Target household growth	2% for growth	Two steps, to 1%
	Operations control—audit score of "4" or "5"	Total checking balance growth	1% for growth 1% for goal	Two steps, to 0.5% Two steps, to 0.5%
		Liability/asset growth	1% for growth 1% for goal	Two steps, to 0.5% Two steps, to 0.5%
	Revenues and margins must meet accountability targets	Revenue growth	2% for growth 2% for goal	Two steps, to 0.5% Two steps, to 0.5%
		Margin growth	2.5% for growth 2.5% for goal	Two steps, to 1% Three steps, to 10%

APPENDIX B
Variable Definitions

Individual Par Score—A rating given to the branch manager for each measure in the balanced scorecard (financial, strategy, customer, control, people, and standards). The par score was expressed as either above par, at par, or below par.

Overall Par Score—A rating given to the branch manager based in part on the six individual par scores for each balanced scorecard category. The par score was expressed as either above par, at par, or below par.

Revenue—Revenue attributed to the branch.

Expense—Expenses matched to branch revenue.

Margin—Revenue minus expenses.

GFS CSI—Score for overall customer satisfaction with overall GFS. This score was expressed as the percentage of survey respondents that rated overall GFS satisfaction in the top two categories on a seven-point scale. A single item was used to measure CSI.

Branch Quality Index—Composite score from survey responses to 20 items. The most heavily weighted item in the branch quality index (45 percent) asked customers to rate ''the overall quality of [the branch's] service against your expectations'' on a five-point scale. The remaining items include the quality of tellers versus expectations (7.5 percent), six additional items concerning tellers (7.5 percent), quality of other branch personnel versus expectations (7.5 percent), six additional items concerning non-teller employees (7.5 percent), quality of automated teller machines (ATMs) versus expectations (7.5 percent), three additional items concerning ATMs (7.5 percent), and one item measuring problem incidence (10 percent). The composite score was translated into an index with a range from 0 to 100.

Household—A group that makes banking decisions as a family or small business.

Retail households—Individual or family branch customers.

B&P households—Business and professional (B&P) branch customers (essentially small businesses).

Premier households—Customers with at least $100,000 combined balances (e.g., checking, saving, investment, etc., balances). Premier customers have access to a variety of services that are not available to other customers.

Household attrition—Number of households (retail, B&P, or premier) that are no longer branch customers.

Lost households—Same as household attrition.

New households—Number of new groups (retail, B&P, or premier) added to the branch.

Cross sales—Revenue generated by GFS service offerings outside of the branch services (i.e., from branch households).

CNR—Customer Net Revenue or the fees and charges paid by branch customers.

AUM—Assets Under Management by branch customer (including checking, saving, etc. balances).

Automated transactions—Banking services completed using technology such as automated teller machines or the Internet.

Salary—Fixed salary paid to branch manager.

Bonus—Additional incentive payment to branch managers above fixed salary. Bonuses are paid on a quarterly basis.

Labor Grade—An internal GFS human resources category assigned to a branch manager based on individual qualifications. The labor grade is used to set the salary and total compensation ranges (or

bands) for an individual branch manager. There are three distinct labor grades used for branch managers.

Time period indicators—Indicator variables that are coded 1 for a specific quarter, and 0 otherwise.

REFERENCES

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

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

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

Banker, R., G. Potter, and D. Srinivasan. 2000. An empirical investigation of an incentive plan that included nonfinancial performance measures. *The Accounting Review* 75 (1): 65–92.

Baron, J., and J. Hershey. 1988. Outcome bias in decision evaluation. *Journal of Personality and Social Psychology* 54 (4): 569–579.

Bellows, R. M. 1954. *Psychology of Personnel in Business and Industry*. Englewood Cliffs, NJ: Prentice Hall.

Bommer, W. H., J. L. Johnson, G. A. Rich, P. M. Podsakoff, and S. B. MacKenzie. 1995. On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology* 48 (3): 587–605.

Blum, M. L., and J. C. Naylor. 1968. *Industrial Psychology*. New York, NY: Harper & Row.

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

Campbell, D., S. Datar, S. Kulp, and V. G. Narayanan. 2002. Using the balanced scorecard as a control system for monitoring and revising corporate strategy. Working paper, Harvard University.

Campbell, J. P. 1990. Modeling the performance prediction problem in industrial and organizational psychology. In *Handbook of Industrial and Organizational Psychology*, Vol. 1, 2nd edition, edited by M. D. Dunnette, and L. M. Hough, 687–732. Palo Alto, CA: Consulting Psychologists Press.

Datar, S., S. Kulp, and R. Lambert. 2001. Balancing performance measures. *Journal of Accounting Research* 39 (1): 75–92.

Feldman, J. M. 1981. Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology* 66 (1): 127–148.

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

Frederickson, J. R., S. A. Pfeffer, and J. Pratt. 1999. Performance evaluation judgments: Effects of prior experience under different performance evaluation schemes and feedback frequencies. *Journal of Accounting Research* 37 (1): 151–165

Ghosh, D., and R. F. Lusch. 2000. Outcome effect, controllability and performance evaluation of managers: Some field evidence from multi-outlet businesses. *Accounting, Organizations and Society* 25 (4–5): 411–425.

Gibbs, M., K. Merchant, W. Van der Stede, and M. Vargus. 2002. Causes and effects of subjectivity in incentives. Working paper, University of Chicago and University of Southern California.

Gjesdal, F. 1981. Accounting for stewardship. *Journal of Accounting Research* 19 (2): 208–231.

Hauser, J. R., D. I. Siemester, and B. Wernerfelt. 1994. Customer satisfaction incentives. *Marketing Science* 13 (3): 327–350.

Hawkins, S. A., and R. Hastie. 1990. Hindsight: Biased judgment of past events after the outcomes are known. *Psychological Bulletin* 107 (3): 311–327.

Hemmer, T. 1996. On the design and choice of ''modern'' management accounting measures. *Journal of Management Accounting Research* 8 (1): 87–116.

Heneman, R. L. 1986. The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. *Personnel Psychology* 39 (4): 811–826.

———, M. L. Moore, and K. N. Wexley. 1987. Performance-rating accuracy: A critical review. *Journal of Business Research* 15 (5): 431–448.

Holmstrom, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10: 74–91.

———, and P. Milgrom. 1991. Multi-task principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization* 7 (1): 24–52.

Hopwood, A. G. 1974. Leadership climate and the use of accounting data in performance evaluation. *The Accounting Review* 49 (1): 485–495.

Ittner, C., D. Larcker, and M. Rajan. 1997. The choice of performance measures in annual bonus contracts. *The Accounting Review* 72 (2): 231–255.

———, and ———. 2002. Determinants of performance measure choices in worker incentive plans. *Journal of Labor Economics* 20 (2): S58–S90.

Kaplan, R. S., and D. P. Norton. 1996. *The Balanced Scorecard: Translating Strategy into Action*. Boston, MA: Harvard Business School Press.

———, and ———. 2001. *The Strategy-Focused Organization: How Balanced Scorecard Companies Thrive in the New Business Environment*. Boston, MA: Harvard Business School Press.

Lambert, R. 2001. Contracting theory and accounting. *Journal of Accounting and Economics* 32 (1–3): 3–87.

Lipe, M. G. 1993. Analyzing the variance investigation decision: The effects of outcomes, mental accounting, and framing. *The Accounting Review* 68 (4): 748–764.

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

———, and ———. 2002. A note on the judgmental effects of the balanced scorecard's information organization. *Accounting, Organizations and Society* 27 (6): 531–540.

Luft, J. and M. Shields. 2001. The effects of financial and nonfinancial performance measures on judgment and decision performance. Working paper, Michigan State University.

MacLeod, W. B. 2001. On optimal contracting with subjective evaluation. Working paper, University of Southern California.

Malina, M. A., and F. H. Selto. 2001 Communicating and controlling strategy: An empirical study of the effectiveness of the balanced scorecard. *Journal of Management Accounting Research* 13 (1): 47–90.

Merchant, K. A. 1989. *Rewarding Results: Motivating Profit Center Managers*. Boston, MA: Harvard University Press.

Mitchell, T., and L. Kalb. 1981. Effects of outcome knowledge and outcome valence on supervisors' evaluations. *Journal of Applied Psychology* 66 (4): 604–612.

Murphy, K., and P. Oyer. 2001. Discretion in executive incentive contracts: Theory and evidence. Working paper, University of Southern California and Stanford University.

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

———, and ———. 1996. Favoritism in organizations. *Journal of Political Economy* 104 (5): 958–978.

Schiff, A. D., and L. R. Hoffman. 1996. An exploration of the use of financial and nonfinancial measures of performance by executives in a service firm. *Behavioral Research in Accounting* 8 (1): 134–151.

Slovic, P., and B. Fischhoff. 1977. On the psychology of experimental surprises. *Journal of Experimental Psychology. Human Perceptions and Performance* 3 (4): 544–551.