THE ACCOUNTING REVIEW
Vol. 75, No. 3
July 2000
pp. 283–298

The Balanced Scorecard: Judgmental Effects of Common and Unique Performance Measures

Marlys Gascho Lipe
University of Oklahoma
Steven E. Salterio
University of Waterloo

ABSTRACT: The balanced scorecard is a new tool that complements traditional measures of business unit performance. The scorecard contains a diverse set of performance measures, including financial performance, customer relations, internal business processes, and learning and growth. Advocates of the balanced scorecard suggest that each unit in the organization should develop and use its own scorecard, choosing measures that capture the unit's business strategy. Our study examines judgmental effects of the balanced scorecard—specifically, how balanced scorecards that include some measures common to multiple units and other measures that are unique to a particular unit affect superiors' evaluations of that unit's performance. Our test shows that only the common measures affect the superiors' evaluations. We discuss the implications of this result for research and practice.

Key Words: *Balanced scorecard, Performance measures, Performance evaluation.*

Data Availability: *Data are available from the second author.*

I. INTRODUCTION

Kaplan and Norton (1992) developed the balanced scorecard (BSC) to complement traditional financial measures of business unit performance. A recent survey estimates 60 percent of Fortune 1000 firms have experimented with the BSC (Silk

We are grateful to the Canadian Academic Accounting Association for project funding and to the University of Alberta's Southam Fellowship for funding for Professor Salterio. Helpful comments were provided by Joan Luft and workshop participants at the University of Alberta, The George Washington University, University of Georgia, Georgia State University, Indiana University, University of Minnesota, University of North Texas, University of Oklahoma, University of Tennessee, University of Texas, and Virginia Polytechnic & State University. We thank Peter Tiessen for help in obtaining subjects.

Submitted November 1998
Accepted February 2000

1998). Adopters include KPMG Peat Marwick, Allstate Insurance, and AT&T (Chow et al. 1997).

The BSC contains a diverse set of performance measures, spanning financial performance, customer relations, internal business processes, and the organization's learning and growth activities (Kaplan and Norton 1992). This large set of measures is designed to capture the firm's desired business strategy (Kaplan and Norton 1993, 1996a) and to include drivers of performance in all areas important to the firm. Use of the BSC should improve managerial decision making by aligning performance measures with the goals and strategies of the firm and the firm's business units.

The BSC is relatively costly to develop[1] so the net benefits gained in adopting the BSC depend on the extent to which it improves managers' decisions. In this study, we explore how managers' cognitive limitations may prevent an organization from fully benefiting from the BSC's information. We examine observable characteristics of the BSC (i.e., measures common to multiple units vs. measures unique to particular units) that may limit managers' ability to fully exploit the information found in a diverse set of performance measures.

Each business unit in the organization develops its own BSC measures to reflect its goals and strategy. While some of these measures are likely to be common across all subsidiaries or units, other measures will be unique to each business unit (Kaplan and Norton 1996b). Judgment and decision-making research suggests that decision makers faced with both common and unique measures may place more weight on common measures than unique measures (Slovic and MacPhillamy 1974). Therefore, managers evaluating multiple subordinate units (i.e., superior managers) may underuse or even ignore the unique measures designed for each unit. Judgmental difficulties in using unique measures may be compounded when the manager who carries out a unit's performance evaluation does not actively participate in developing that unit's scorecard and, consequently, may not appreciate the significance of the unique measures. Underuse of unique measures reduces the potential benefits of the BSC because the unique measures are important in capturing the unit's business strategy.

To investigate whether common measures dominate BSC-based evaluations of subordinate units, we develop an experiment where M.B.A. students evaluate two divisions of a clothing firm. The two divisions sell to different markets and have different business strategies. They have balanced scorecards with some common and some unique measures. We manipulate the performance of the two divisions (relative to their targets) on their common and unique measures in a crossed design. Division one may outperform (or underperform) division two on common measures, and division one may also outperform (or underperform) division two on unique measures. Our results show that the experimental participants evaluate the divisions based solely on the common measures. Performance on unique measures has no effect on the evaluation judgments.

Our study is the first to document a cognitive difficulty in using the BSC, and our result that superior managers appear to disregard unique measures in performance evaluations has implications for managers and firms using the BSC. If unique measures do not affect subordinates' *ex post* performance evaluations, then the subordinate manager is unlikely to use unique measures in *ex ante* decision making (Holmstrom and Milgrom 1991). In addition, Kaplan and Norton (1996b) note that measures that are common across units

[1] Management time is one element of this cost. Kaplan and Norton (1996b) estimate that development could take as little as 16 weeks, but Chow et al. (1997) and Kaplan and Norton (1996a) indicate development and implementation can require a significant time investment for two years or more.

often tend to be lagging and financial indicators of performance. In contrast, unique measures are more often leading and nonfinancial measures. Consequently, our results suggest that managers may pay insufficient attention to leading and nonfinancial measures. This defeats the purpose of implementing the BSC, which is to expand the set of measures that managers use in decision making (Kaplan and Norton 1996b). If the unique measures on the scorecard do not affect managers' decisions, firms will not reap the expected benefits of BSC adoption.

The remainder of the paper is organized as follows. The next section describes the BSC and its use, as envisioned by Kaplan and Norton (1996b). In Section III we review the judgment and decision-making literature applicable to the study of the BSC and develop our research hypothesis. Sections IV and V describe the experimental method and results, and the final section discusses the implications and limitations of the study.

II. THE BALANCED SCORECARD

Implementation

Kaplan and Norton's (1996b) best-selling book provides a blueprint for organizations interested in implementing a BSC. The BSC is an integrated set of leading and lagging performance measures designed to capture the organization's strategy. Kaplan and Norton (1996a, 1996b) identify four major steps in implementing a BSC: (1) clarifying and translating the vision and strategy, (2) communicating and linking, (3) planning and target setting, and (4) strategic feedback and learning.

The first step, clarifying and translating the vision and strategy, is generally accomplished by a team of upper management, although Kaplan and Norton (1996b) indicate that this has been successfully accomplished by a single senior executive. The purpose of this phase is to develop an understanding of the firm's mission and strategy for obtaining its goals. Since mission statements are often vague, management must translate the mission into specific objectives and then develop a strategy that will use the firm's strengths to meet the objectives In doing so, management should develop a set of measures that captures this strategy. This will become the organization's BSC.

After the firm's BSC has been developed, each strategic business unit determines measures for its own scorecard as part of the communicating and linking step. Unit managers consider both the overall organizational objectives and strategy and their own units' business strategy. The measures chosen for the unit's BSC should describe what the unit must do to accomplish its strategy, which will in turn help the organization accomplish its objectives. Kaplan and Norton (1993, 135) state that the measures on the unit's scorecard should be specifically designed to fit the unit's "mission, strategy, technology, and culture." For example, when Mobil Corp.'s Americas Marketing and Refining division showed its unit leaders the division's BSC, the leaders were told: "Develop a set of strategic objectives and measures that represent the direction of your business. Here's a copy of the division's scorecard, but don't worry about making yours a carbon copy" (McWilliams 1996, 18). Similarly, in discussing the directions given to divisional managers at FMC Corporation, executive vice president Larry Brady stated, "We definitely wanted the division managers to perform their own strategic analysis and to develop their own measures" (Kaplan and Norton 1993, 145).

In the normal course of a BSC implementation, each unit manager (and his or her team) develops that unit's scorecard, while higher-level managers approve the scorecards and use them for evaluation and further decision making. Chow et al. (1997) indicate that even small businesses develop multiple scorecards, each tailored to the strategy and goals of a

specific subunit. Thus, the second step of BSC implementation requires many people in units throughout the organization to develop scorecards for their particular segments of the business.

In the remaining steps of BSC implementation, managers set targets and budgets (step three), and over time, receive feedback on the strategies of the business units and the firm by evaluating performance relative to the scorecard measures (step four). These activities parallel those performed in non-BSC firms so we do not discuss them in detail. We now describe the types of performance measures included in the BSC.

Categories of Measures

The BSC should include measures of financial performance, customer relations, internal business processes, and organizational learning and growth (Kaplan and Norton 1996b). As discussed above, some of the specific measures chosen for each individual business unit in the organization will likely differ from those of other units, because the measures should be tailored to each unit's specific goals and strategies.

Financial measures can include traditional measures such as return on assets and net income, but Kaplan and Norton (1996b) advocate measures that are particularly relevant to the business unit (e.g., revenues per employee for a sales unit or research and development expense/sales for a pharmaceutical division). Measures related to customers include results of customer surveys, sales from repeat customers, and customer profitability. Internal business process measures relate specifically to the operational processes of the business unit. For example, a petroleum distributor may measure dealer acceptance of new distributor programs and dealer quality (Kaplan and Norton 1996b, 111–113). The final set of performance measures relates to learning and growth and they are often the most difficult to select. Kaplan and Norton (1996b, 127) suggest measures of employee capabilities, information systems capabilities, and employee motivation and empowerment.

Linking the BSC to Performance Evaluation and Compensation

Kaplan and Norton (1996b, 217–223) indicate that it is problematic to ask managers to focus on BSC measures if managers' compensation and evaluation are based on traditional financial measures. However, Kaplan and Norton (1996b) do not provide specific recommendations regarding how to link the BSC to compensation.[2] In light of Kaplan and Norton's (1996b) reticence on the link between the BSC and compensation, our experiment investigates the use of the BSC for performance measurement and evaluation, rather than for compensation. Kaplan and Norton (1996b, 1996a) indicate that managers will use the BSC for performance evaluation during operational reviews. Operational reviews of performance using the BSC are similar to those in non-BSC firms (i.e., a manager evaluates subordinate units and unit managers) except that the BSC provides the evaluating manager with a broader set of measures (including nonfinancial measures) to use in assessing subordinates' performance. Thus, we consider the effect of this broader set of measures in the context of an organization's performance review and evaluation.

The BSC is a relatively complex and costly measurement system. It is important to understand how cognition affects use of the BSC in order to understand how managers'

[2] Chow et al. (1997) and McWilliams (1996) argue that it is important to link the BSC measures with compensation, although the firm may want to use the scorecard for some time before establishing the link. Major firms including CIGNA Insurance and Sears Roebuck, have reportedly linked employee compensation with corporate or business unit scorecard performance (Epstein and Manzoni 1997; Rucci et al. 1998).

cognitive capabilities and characteristics may limit the BSC's potential benefits. The next section reviews judgment and decision-making research relevant to assessing managers' abilities to process and use the BSC information.

III. JUDGMENT AND DECISION-MAKING RESEARCH

The Use of Common and Unique Information

Kaplan and Norton (1996b) argue that an important strength of the BSC is that each business unit in the organization will have its own scorecard specifically tailored to that unit. Yet Kaplan and Norton (1996b, 149) also note that all balanced scorecards are likely to use "certain generic measures." Thus, units at the same organizational level will have some common measures in addition to others that are unique to their business and strategy. Since operational reviews of performance typically occur at regular intervals (e.g., quarterly), the superior manager makes evaluations of multiple subordinate units (and their managers) within a short time frame, based on both common and unique measures.

A classic judgment and decision-making study suggests that people use common and unique information differently. Slovic and MacPhillamy's (1974) undergraduate volunteer participants judged which of two college students had the higher freshman GPA. The experimental subjects based their judgments on numerical information regarding the students' English skills, quantitative aptitude, and need for achievement. Participants saw some information that was common to the two college students and other unique information. For example, the material could provide scores on English skills for both students, a quantitative aptitude score for student one, and a need for achievement score for student two. The participants saw such information for pairs of students and judged which student would have the higher freshman GPA and the magnitude of the difference. Slovic and MacPhillamy (1974) found that participants weighted the common measures more heavily than unique measures for both the judgment and the choice. Monetary incentives and feedback did not eliminate this differential weighting. Further, the effect did not result from differential diagnosticity of the common and unique items; another group of participants who predicted the GPAs when each student's information was presented individually (i.e., no comparison case was given) did not weight the information items differentially.

Slovic and MacPhillamy (1974) argue that common information has a greater impact because it is easier to use in comparing the candidates. This suggests that evaluators in a BSC firm, faced with common and unique measures across business units, may concentrate on the common measures to simplify their judgment task (Payne et al. 1993). However, there are several reasons why this emphasis on common measures may not occur in the context of BSC evaluations.

Although Slovic and MacPhillamy (1974) find that common measures dominate unique measures, the dominance is not very large. Slovic and MacPhillamy (1974, 184) report that, for their comparative judgments, the average weight placed on common measures is 0.395 and on unique measures is 0.342. The authors also note a potential confound in their tests (Slovic and MacPhillamy 1974, 189). In many of the pairs of hypothetical students, a decision strategy of weighting each piece of information equally would lead to the same choice as a decision strategy based on only the common measures. Thus, although Slovic and MacPhillamy (1974) conclude that common measures are dominant for comparative judgments, the magnitude of dominance is small and the results cannot be interpreted unambiguously.

Furthermore, Slovic and MacPhillamy's (1974) results may not obtain in the BSC context if superior managers evaluate each business unit and unit manager *individually*.

Each of the units has a scorecard with measures tailored to its strategy, along with targets for those measures that are also tailored to the unit's business environment. The superior manager is *not* asked to directly compare the units or subordinate managers in making evaluation judgments. In Slovic and MacPhillamy's (1974) study common measures dominated only for comparative judgments. Thus, if the superior managers judge each business unit individually, they may not weight the common measures more heavily.

Further, in Slovic and MacPhillamy's (1974) study, participants did not have any *particular* knowledge regarding the importance of the three items of information (e.g., English skills). In contrast, the superior manager in a BSC firm knows the subordinate business unit's strategy. This should help him or her appreciate the importance of the unique measures in evaluating the success of the strategy's implementation. Thus, in the BSC situation, the evaluator's knowledge may counteract the tendency to focus on common measures.

In sum, although a classic judgment and decision-making study showed that common items dominated unique items in a grade prediction task, it is unclear whether the same effect will arise in BSC evaluations. While we believe that strategically adaptive decision making is often automatic and unconscious and that judgments are inherently comparative (Hsee 1996, 1998), the weighting of common and unique measures used on balanced scorecards is an untested empirical issue. We therefore posit the following hypothesis stated in null form.

H_N: Performance evaluations using the balanced scorecard will be affected by *both* unique measures and common measures.

The next section describes the experimental test of the hypothesis.

IV. METHOD

Overview of Experiment

Participants in the experiment read a case asking them to take the role of a senior executive of WCS Incorporated, a firm specializing in women's apparel.[3] The case materials focus on WCS's two largest divisions. The case indicates that the firm's chief financial officer attended a Harvard Business School symposium on the balanced scorecard. Further, it describes the BSC concept and lists the four categories of measures. Participants learned that WCS decided to implement the BSC. The case quotes WCS's mission statement,[4] introduces the managers of the two business units (divisions), describes the strategies of the divisions, and presents a balanced scorecard for each division. Table 1 presents one division's balanced scorecard. Participants acted as a WCS senior executive (superior), making the following judgment for each (subordinate) unit's manager.[5]

---

[3] The authors (one of whom has prior experience with apparel retailing) developed the case following Kaplan and Norton's (1996b) Kenyon Stores example.

[4] The mission statement says, "We will be an outstanding apparel supplier in each of the specialty niches served by WCS."

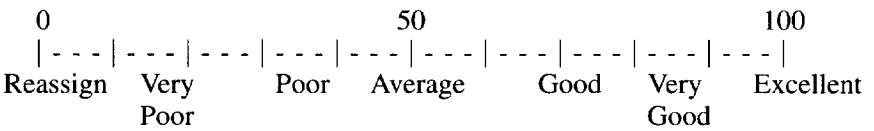
[5] The following qualitative descriptors were listed below the scale:
Excellent: far beyond expectations, manager excels
Very good: considerably above expectations
Good: somewhat above expectations
Average: meets expectations
Poor: somewhat below expectations, needs some improvement
Very poor: considerably below expectations, needs considerable improvement
Reassign: sufficient improvement unlikely

TABLE 1
The Balanced Scorecard for RadWear Division Employed in the Experiment[a]

Measure	*Target*	*Actual*	*Percent Better than Target*
Financial:			
1. Return on sales	24%	26%	8.33
2. New store sales	30%	32.5%	8.33
3. Sales growth	35%	38%	8.57
4. Market share relative to retail space	$80	$86.85	8.56
Customer-Related:			
1. Mystery shopper program rating	85	96	12.94
2. Repeat sales	30%	34%	13.33
3. Returns by customers as percentage of sales	12%	11.6%	3.33
4. Customer satisfaction rating	92%	95%	3.26
Internal Business Processes:			
1. Returns to suppliers	6%	5%	16.67
3. Average major brand names/store	32	37	15.63
3. Average markdowns	16%	13.5%	15.63
4. Sales from new market leaders	25%	29%	16.00
Learning and Growth:			
1. Average tenure of sales personnel	1.4	1.6	14.29
2. Hours of employee training/employee	15	17	13.33
3. Stores computerizing	85%	90%	5.88
4. Employee suggestions/employee	3.3	3.5	6.06

[a] A similar scorecard was prepared for WorkWear, the other division in the experiment.

Indicate your initial performance evaluation for this manager by placing an 'X' somewhere on the scale below.



This task follows Kaplan and Norton (1996b, 1996a) in that each unit develops its own scorecard and a superior manager evaluates the unit relative to this scorecard.

After evaluating the unit managers, participants completed a debriefing questionnaire. This asked for demographic information, provided manipulation checks (discussed further in the results below), and gathered data regarding task difficulty, realism, and understandability.

The two divisions described are RadWear, a retail division specializing in clothing for the urban teenager, and WorkWear, a division that sells business uniforms through direct sales contacts with business clients. The case describes the strategies of the two divisions as follows.

> RadWear's management determined that its growth must take place through an aggressive strategy of opening new stores. RadWear also determined that it must

increase the number of brands offered to keep the attention and capture the clothing dollars of its teenage customers. RadWear concluded that its competition radius is fairly small due to the low mobility of young teens.

Although WCS has historically focused on women's clothing, WorkWear's management decided to grow its sales by including a few basic uniforms for men. It is expected that this will make WorkWear a more attractive supplier for businesses that want to purchase uniforms from a single supplier. WorkWear also decided to print a catalog so that clients could place some orders without a direct sales visit, particularly for repeat or replacement orders; this should help to retain some sales which might otherwise be lost due to time considerations.

The performance measures for each division are appropriate for that division and capture these strategies.

Subjects

Fifty-eight first year M.B.A. students served as experimental participants. The students had, on average, more than five years of work experience and 63 percent were male.

Design and Procedure

The experiment employs a 2 × 2 between subjects (Ss) design, in conjunction with a 2-level within-Ss factor (i.e., the complete design is 2 × 2 × 2). The first independent factor indicates the particular pattern of performance for the two business units based on their common measures. Thus, RadWear could perform better on the common measures than WorkWear (COM-Rad) or WorkWear could outperform RadWear on the common measures (COM-Work). Similarly, the second factor is the particular pattern of performance for RadWear and WorkWear based on their unique measures. So RadWear could perform better on its unique measures than does WorkWear on its unique factors (UNIQ-Rad) or vice versa (UNIQ-Work). Each subject evaluated both divisional managers, thus division (RadWear, WorkWear) is the within-Ss factor.

We developed separate 16-measure balanced scorecards for WCS's two divisions. The scorecards contained four performance measures in each of the BSC categories. Each category included two common measures (i.e., used for both divisions) and two unique measures (i.e., designed only for RadWear or only for WorkWear). Table 2 lists the performance measures. For all measures, each division performed better than its target. The percentage above target, however, varied in the design as indicated above.

We chose all performance data so that common and unique items had the same excess performance. For example, in COM-Rad the first common financial measure (i.e., Return on Sales) was 8.33 percent above target for RadWear and 4.17 percent above target for WorkWear. Similarly, in UNIQ-Rad, the first unique financial measure was 8.33 percent above target for RadWear and 4.25 percent above target for WorkWear. Although the exact percentages varied slightly due to rounding, we counterbalanced and controlled even these small variations. The sum of excess performance (percentage above target) for common measures was 85.18 for the better division and 51.99 for the worse division (a difference of 33.19). The sum of excess performance for unique measures ranged from 84.96 to 85.08 for the better division and 51.20 to 51.95 for the worse division (a difference of 33.01 or 33.88). The slight variation for the unique measures was due to the different units or bases for the measures related to RadWear vs. WorkWear. (This strict control over performance levels eliminates the potential confound present in Slovic and MacPhillamy [1974], as

TABLE 2
Common and Unique Performance Measures for RadWear and WorkWear Divisions' Balanced Scorecards

Type[a]	*Measure*
Financial Measures:	
C	Return on sales
C	Sales growth
U-Rad	New store sales
U-Rad	Market share relative to retail space
U-Work	Revenues per sales visit
U-Work	Catalog profits
Customer-Related Measures:	
C	Repeat sales
C	Customer satisfaction rating
U-Rad	Mystery shopper program rating
U-Rad	Returns by customers as a percent of sales
U-Work	Captured customers
U-Work	Referrals
Internal Business Process Measures:	
C	Returns to suppliers
C	Average markdowns
U-Rad	Average major brand names per store
U-Rad	Sales from new market leaders
U-Work	Orders filled within one week
U-Work	Catalog orders filled with errors
Learning and Growth Measures:	
C	Hours of employee training per employee
C	Employee suggestions per employee
U-Rad	Average tenure of sales personnel
U-Rad	Stores computerizing
U-Work	Percent sales managers with M.B.A. degrees
U-Work	Database certification of clerks

[a] C indicates a common measure. U-Rad is a unique measure for RadWear, a teen-wear retail division. U-Work is a unique measure for WorkWear, a uniform division that sells through catalogs and sales calls. Participants received a brief written description of the calculation of each measure.

discussed previously). Table 1 shows that the percent "better than target," calculated to the second digit, appears as a column in the scorecards presented to the participants.

Pilot Testing the Instrument

The experiment is designed to compare how subjects respond to common vs. unique measures, so participants must believe these two sets of measures have similar validity. Unfortunately, we could not use a specific item (e.g., catalog profits) as a common measure for some subjects and a unique measure for others. This is not possible because measures for each division must link to that division's strategy (and catalog profits would be inappropriate for RadWear, for example). Instead, we performed several tests to determine whether participants perceived the two sets of measures similarly.

First, we tested whether the groups of unique measures chosen for RadWear and WorkWear were similar in their relation to the BSC categories. A group of 14 M.B.A. students (who did not participate in the experiment) rated how typical each measure was of its category. The rating scale ranged from 0 "not at all typical" to 10 "very typical." The average rating was 7.13 and the typicality of RadWear's unique measures did not differ from WorkWear's unique measures (means of 7.00 and 7.25, respectively; $t = 0.64$, $p > 0.5$). This indicates that the unique measures we chose for the two divisions were similarly representative of their BSC categories. Thus, subjects should not differ in their use of unique measures simply due to a misunderstanding regarding these measures for one division.

Second and more important, we tested whether common and unique measures were equally relevant to the performance evaluation judgments we asked the subjects to make. We listed the 16 measures for each division and asked 18 masters of accounting students (who did not participate in the experiment) to rate each of these measures as to "its relevance for evaluating the performance of the division manager," using a scale from 1 "low relevance" to 10 "high relevance." There were no differences between common and unique measures' decision relevance in total (average relevance scores of 7.17 and 7.02, respectively, $t = 0.84$) or for RadWear (7.07 vs. 7.28, $t = 1.44$) or WorkWear (7.29 vs. 7.28, $t = 0.08$) individually. Thus, any difference in subjects' use of common and unique measures is not likely attributable to differences in perceived relevance of these measures.

Finally, to broaden our checks beyond our pre-test subjects we examined a study by Dempsey et al. (1997). Dempsey et al. (1997) asked financial analysts to rate a set of strategic measures as to their frequency of use and predictive value for the analysts; ten of our measures (five common and five unique) appeared in their list. Using Dempsey et al.'s (1997) data, there were no significant differences between our common and unique measures in either analysts' frequency of use ratings (means on a scale of 1 to 5 of 2.64 and 2.63, respectively, $t = 0.02$) or predictive values (2.83 and 2.80, respectively, $t = 0.12$).

In summary, our pilot tests indicate that our unique measures for the two divisions are similarly representative of their BSC categories. The tests also indicate that although we could not use exactly the same items as both common and unique measures, our common and unique measures are similarly relevant to divisional performance judgments and similarly important to financial analysts in evaluating firms. Further, as noted above, we carefully controlled the particular "percent above target" values for the common and unique measures so that differences in performance would not confound our results.

Dependent Measure

All subjects evaluated the manager of RadWear as well as the manager of WorkWear. We want to determine whether performance on common and unique measures affects subjects' evaluations of the division managers. If common measures affect these evaluations, we will observe an interaction of division and common measures. If unique measures affect these evaluations, we will find an interaction of division and unique measures.

V. RESULTS

A manipulation check shows the participants recognize that the two divisions employ different performance measures ($p < 0.01$). Further manipulation checks show that participants agree the two divisions sell to different markets ($p < 0.01$) and that it is appropriate for the divisions to employ different performance measures ($p < 0.01$). In addition, there are no differences across experimental treatments in ease of understanding, case difficulty,

or case realism (all p-values > 0.10). The manipulation check results do not vary across experimental treatments.

We used a 2 × 2 × 2 repeated-measures ANOVA to test our hypothesis. The results appear in Panel A of Table 3. The only statistically significant effect is the interaction of common measures and division ($F = 30.69$, $df = 1,54$, $p < 0.01$) indicating that the pattern of performance on common measures affects the managers' evaluations, while the pattern for unique measures does not.[6] Panel B of Table 3 indicates that when common measures favor RadWear, the superior manager evaluates RadWear's manager 5.97 points higher than WorkWear's manager. Similarly, when common measures favor WorkWear, WorkWear's manager is evaluated 7.17 points higher than RadWear's manager. In contrast, when unique measures favor RadWear (WorkWear), there is no significant difference in the evaluations of the managers, a mean difference of 0.64 (1.77). We also used regression analysis, regressing the differences in managerial performance evaluations (RadWear's evaluation minus WorkWear's evaluation) on the relative performance on common measures (coded "1" when RadWear outperformed WorkWear and "0" otherwise) and the relative performance on unique measures (analogously coded). The common measures have a significantly positive slope coefficient of 10.87 ($t = 3.28$, $p < 0.01$) while unique items' coefficient of 0.08 is insignificant ($t = 0.02$, $p > 0.10$).[7] Thus, the results suggest that Slovic and MacPhillamy's (1974) finding of a natural simplifying strategy whereby common measures dominate unique measures also applies in a BSC context.

VI. IMPLICATIONS, LIMITATIONS, AND RESEARCH ISSUES

This study shows that unique measures in a business unit's BSC may be underweighted in performance evaluation. This section describes implications of this result, acknowledges some limitations of the study, and discusses special research challenges arising in studies of real-world accounting and management phenomena.

Implications of the Results

Our evidence concerning the disregard or underuse of unique measures in evaluating business unit performance has two major implications. First, our evidence that unique measures are disregarded in the *ex post* performance evaluation of a business unit's manager has significant implications for the unit manager's *ex ante* decision-making strategy. Holmstrom and Milgrom (1991) show analytically that agents' decisions are affected by items that are included in their performance evaluation and compensation. They also show that items not included in evaluation and compensation of an agent will have little effect on the agent's decisions. Psychological research (e.g., McNamara and Fisch 1964) has shown this same result experimentally. Thus, our results suggest that common measures that drive the unit managers' evaluations will have more effect on unit managers' decisions than will the unique measures that are not used in the evaluations.[8]

---

[6] Tests of ANOVA model assumptions indicate no problems except for nonnormality of the error terms. This is not unusual for a study including repeated measures. Although the F-tests are quite robust to this nonnormality, we also ran an additional ANOVA using the difference between the performance evaluations of the managers of the two divisions as the dependent variable (RadWear's evaluation minus WorkWear's evaluation). This analysis meets all ANOVA assumptions and the results corroborate the reported results (i.e., only common measures have a significant effect on the difference in the managers' evaluations, $F = 30.69$, $p < 0.01$).

[7] We also included the interaction of the common and unique measures; the effect was statistically insignificant ($t = 0.99$, $p > 0.3$). The regression's adjusted R^2 is 0.34.

[8] Davila and Simons' (1997) case study on performance evaluation at Citibank indicates senior management's concern regarding the effect of performance evaluation on future unit decision making. In this case, the unit manager performed poorly in one BSC category. The senior manager worries that "[i]f the performance evaluation team gave [the evaluatee] an "above par" [rating] people could think the division was not serious about its nonfinancial measures" (Davila and Simon 1997, 4).

TABLE 3
Experimental Results for Managers' Performance Evaluations

Panel A: Results of a 2 × 2 × 2 Repeated Measures ANOVA of Evaluations of the Performance of RadWear and WorkWear Divisions' Managers

Variable	*df*	*SS*	*MS*	*F*	*p*
Between Subjects:					
Common	1	143.86	143.86	0.54	0.47
Unique	1	57.16	57.16	0.21	0.65
Common × Unique	1	434.13	434.13	1.63	0.21
Error	54	14,402.36	266.71		
Within Subjects:					
Division	1	9.15	9.15	0.22	0.64
Division × Common	1	1,265.38	1,265.38	30.69	0.00
Division × Unique	1	42.04	42.04	1.02	0.32
Division × Common × Unique	1	40.07	40.07	0.97	0.33
Error	54	2,226.44	41.23		

Panel B: Evaluations of the Performance of RadWear and WorkWear Divisions' Managers[a]

Measures		*Favor RadWear*	*Favor WorkWear*
Common	RadWear	74.21[b]	70.00
		(11.08)	(12.86)
	WorkWear	68.24	77.17
		(14.26)	(11.09)
Difference: RadWear – WorkWear		5.97	−7.17
Unique	RadWear	72.00	72.20
		(11.02)	(13.19)
	WorkWear	71.36	73.97
		(11.70)	(14.97)
Difference: RadWear – WorkWear		0.64	−1.77

[a] Evaluations made on a 101-point scale, with 0 labeled "Reassign" and 100 labeled "Excellent."

[b] Panel values are mean (standard deviation). *Common* measures appear on both divisions' balanced scorecards, *Unique* measures appear on only one division's balanced scorecard. *Favor Radwear* indicates the measures were higher for the RadWear division than the WorkWear division. *Favor WorkWear* indicates the measures were higher for the WorkWear division than the RadWear Division.

Second, Kaplan and Norton (1996b) note that lagging measures are often rather generic (i.e., applicable to many units), while leading measures are more likely to be customized for each business unit. Thus, evaluators who focus on common measures may largely overlook or disregard leading measures. Further, since financial measures are often common across business units, the unique nonfinancial measures may receive less attention. Underweighting nonfinancial and leading measures undermines the goals of the BSC, which was expressly designed to incorporate such measures into managerial thought and decision making (Kaplan and Norton 1996a, 1996b).

Limitations of the Study

Our experimental design has limitations. First, our experimental participants were not involved in development of the units' scorecards. Thus, we are unable to investigate the effect of such involvement, although greater involvement might increase reliance on all the BSC measures, including the unique measures. However, our experiment captures the hierarchical nature of business organizations where higher-level managers evaluate lower-level unit managers, and these higher-level managers are not directly involved in unit BSC development. A second limitation is that our participants may have been novices in the use of the BSC, and they did not necessarily have business experience in the retail and apparel sector from which we developed our case materials. Third, although our experiment carefully controlled performance on common and unique measures, it is possible that the perceived performance relative to the targets differed for the various groups of measures. While our pilot tests and manipulation checks did not reveal any such problems, this possibility cannot be ruled out.

Research Obstacles

Recent and widespread interest in the BSC motivated this study. Our knowledge of the judgment and decision-making literature led us to believe that use of this tool might be impaired by managers' cognitive limitations, and that these limitations would lead to predictable judgment effects. However, this study was primarily motivated by a desire to better understand a real-world phenomenon. We offer some observations regarding special difficulties in conducting research motivated by such phenomena.

Ideally, empirical accounting research should be conducted in the context of relevant theories and models. This can be particularly difficult for studies focusing on topical accounting and management issues. Analytical models and general theories require simplifying assumptions to maintain tractability. The realism of these assumptions determines whether the models and their predictions can be applied directly to the phenomenon of interest. Even when these models cannot be applied directly, researchers can still use the models to (1) suggest variables to investigate through manipulation or direct measurement, (2) sensitize the researcher to variables that should be covariates or controls, and (3) indicate particularly problematic variables that the researcher cannot control but must consider in drawing conclusions from the study. For example, Banker and Datar's (1989) model indicates the importance of controlling the decision relevance of common and unique measures in our task. This insight prompted us to use pilot tests to ensure that participants perceived common and unique measures as equally decision-relevant.

The inability to directly apply analytical models to many real-world situations means that a normative criterion against which to evaluate judgments and decisions will often be absent. For example, there is no normative model for performance evaluation scores, so we cannot assess the accuracy of our participants' evaluations in this study. Although the frequent absence of a normative criterion can be a disadvantage relative to studies that are motivated by, and designed around, a test of a normative theory, researchers can deal with this challenge in at least two ways. In our case, we designed our study to test for *differences* in judgment across levels of common and unique measures. Knowing the correct evaluation scores is not necessary to test and show that common measures have more judgmental impact than unique measures, contrary to the spirit of the BSC. Illustrating another approach, Cloyd and Spilker (1999) and Libby and Libby (1989) use panels of experts to determine judgment criterion for realistic tax and audit tasks, respectively. Thus, the lack of a normative model need not preclude investigation of a research question, although it presents challenges to the researcher.

Determining the scope of a study motivated by a real-world phenomenon is also challenging. In research designed to test a specific theory, the scope of the study is determined by that theory. The number of factors, the number of tests, issues that are central and those peripheral to the study are all determined by the theory. In contrast, the scope of a study motivated by a real-world phenomenon is less determinate; it is often unclear how many factors must be included or which factors are most important for study. Although models and theories suggest potentially important factors, the choice of models and, consequently, variables is a matter of researcher knowledge and interest. For example, we base our study on prior psychological work on the use of common and unique information items. Future BSC research may investigate other attributes, such as precision or sensitivity of the measures (Banker and Datar 1989), or could explore processing issues such as information overload (Schick et al. 1990) or information organization effects. However, it remains unclear how many such issues a single study could feasibly address.

Investigating a real-world phenomenon leads to a natural desire to report how firms are actually doing business and how many firms are using the methods of interest. Such data can be very difficult to obtain, particularly as firms make the transition to the new methods or, as in the case of the BSC, when firms consider the information proprietary and sensitive (Kaplan and Norton 1996b, 148). We would like to report how many firms use the BSC and we would like to provide data regarding the prevalence and types of common and unique measures, but such information is not readily available.

Absent relevant archival data, researchers may use experimental tests to study real-world phenomena. However, it can be difficult to identify appropriate participants for the experiments. We chose to use graduate business students who were novices in the use of the BSC. Other potential subject groups include employees of a particular BSC firm, employees in a cross-section of BSC firms, or people who have been trained in the use of a BSC. These different subject pools have different advantages and disadvantages, and theory does not suggest an optimal choice. In a study of the use of common vs. unique measures, focusing on employees of a particular BSC firm as they use their own actual scorecards would compromise experimental control. Specifically, performance on common measures would need to differ substantially from performance on unique measures in order to determine the effects of the two types of measures. Using employees from a cross-section of BSC firms means that most or all participants would be presented with scorecards different from those used in their own firms and units. In this case, participants' own experiences with specific measures may differ from the situation presented in the experimental materials (e.g., catalog profits is a unique measure in our experiment, but it could be a common measure in some firms). This might lead a participant to weight a measure based on his or her experience with the measure (e.g., high weight for a measure that is common in the participant's firm) rather than responding to the manipulated level of the measure (e.g., unique in the experimental instrument) so that test results may not capture the effects of manipulated variables. Finally, training experimental participants to use a scorecard can lead to significant demand effects. For example, if we trained our student-subjects to weight BSC items in a particular way or to use a particular strategy for interpreting the scorecards (e.g., using a causal diagram linking the measures to one another), our tests of the use of measures would largely reflect the success of our training program. Thus, although it is tempting to use subjects with BSC experience, this presents difficulties for a controlled and generalizable study.

Conclusion

Interest in the balanced scorecard continues to grow. However, proponents, adopters, and researchers know very little about the human information processing demands of the

BSC. Judgment and decision-making research indicates that humans use simplifying decision strategies that are affected by task characteristics. We investigated the effect on performance judgments of the BSC's inclusion of diverse measures. Our experimental participants succumbed to the simplifying strategy of using only common measures in evaluating multiple managers.[9] If unique measures are underweighted in *ex post* evaluations of the business unit and its manager, these measures are likely to receive little *ex ante* weight in the unit's decisions (Holmstrom and Milgrom 1991). This focus on common measures undermines one of the major espoused benefits of the BSC, namely, that each business unit will have and use a scorecard that uniquely captures its business strategy.

[9] Interestingly, accounting practitioner journals and management journals recommend remedies for assimilating common and unique performance measures. Lee (1992) and Nourayi and Daroca (1996) both recommend that evaluators convert unique measures into common measurement scales via indices. While such aggregation can alleviate cognitive difficulties of working with unique measures, it sacrifices the BSC characteristic of using specific individual measures clearly tied to the business unit's strategic goals.

REFERENCES

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

Chow, C. W., K. M. Haddad, and J. E. Williamson. 1997. Applying the balanced scorecard to small companies. *Management Accounting* 79 (2): 21–27.

Cloyd, B., and B. Spilker. 1999. The influence of client preferences on tax professionals' search for judicial precedents, subsequent judgments and recommendations. *The Accounting Review* 74 (3): 299–322.

Davila, A., and R. Simons. 1997. *Citibank: Performance Evaluation.* Harvard Business School Case 9-198-048 (revised April 20, 1998). Boston, MA: Harvard Business School Publishing.

Dempsey, S., J. F. Gatti, D. J. Grinnell, and W. L. Cats-Baril. 1997. The use of strategic performance variables as leading indicators in financial analysts' forecasts. *The Journal of Financial Statement Analysis* 2 (4): 61–79.

Epstein, M., and J. Manzoni. 1997. The balanced scorecard and Tableau de Bord: Translating strategy into action. *Management Accounting* 79 (2): 28–36.

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

Hsee, C. 1996. The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavioral and Human Decision Processes* 67 (3): 247–257.

———. 1998. Less is better: When low-value options are valued more highly than high-value options. *Journal of Behavioral Decision Making* 11 (2): 107–121.

Kaplan, R., and D. Norton. 1992. The balanced scorecard—Measures that drive performance. *Harvard Business Review* 70 (1): 71–79.

———, and ———. 1993. Putting the balanced scorecard to work. *Harvard Business Review* 71 (5): 134–147.

———, and ———. 1996a. Using the balanced scorecard as a strategic management system. *Harvard Business Review* 74 (1): 75–85.

———, and ———. 1996b. *The Balanced Scorecard.* Boston, MA: Harvard Business School Press.

Lee, J. Y. 1992. How to make financial and nonfinancial data add up. *Journal of Accountancy* 174 (3): 62–66.

Libby, R., and P. A. Libby. 1989. Expert measurement and mechanical combination in control reliance decisions. *The Accounting Review* 64 (4): 729–747.

McNamara, H., and R. Fisch. 1964. Effect of high and low motivation on two aspects of attention. *Perceptual and Motor Skills* 19: 571–578.

McWilliams, B. 1996. The measure of success. *Across the Board* 33 (2): 16–20.

Nourayi, M. M., and F. P. Daroca. 1996. Performance evaluation and measurement issues. *Journal of Managerial Issues* 8 (2): 206–217.

Payne, J., J. Bettman, and E. Johnson. 1993. *The Adaptive Decision Maker*. Cambridge, U.K.: Cambridge University Press.

Rucci, A., S. Kirn, and R. Quinn. 1998. The employee-customer-profit chain at Sears. *Harvard Business Review* 76 (1): 82–97.

Schick, A., L. Gordon, and S. Haka. 1990. Information overload: A temporal approach. *Accounting, Organizations and Society* 15 (3): 199–220.

Silk, S. 1998. Automating the balanced scorecard. *Management Accounting* (May): 38–44.

Slovic, P., and D. MacPhillamy. 1974. Dimensional commensurability and cue utilization in comparative judgment. *Organizational Behavior and Human Performance* 11: 172–194.