PERGAMON

Accounting, Organizations and Society 27 (2002) 531–540

Accounting, Organizations and Society

www.elsevier.com/locate/aos

A note on the judgmental effects of the balanced scorecard's information organization

Marlys Gascho Lipe[a,*], Steven Salterio[b]

[a]*University of Oklahoma, Price College of Business, Norman, OK 73019, USA*
[b]*University of Waterloo, School of Accountancy, Waterloo, Ontario, Canada N2L 3G1*

Abstract

We examine judgmental effects of the balanced scorecard's organization. The balanced scorecard contains a large number of performance measures divided into four categories. We examine whether the scorecard's organization results in managerial performance evaluation judgments consistent with a recognition of the potential relations (i.e. nonindependence) of measures within a category. Supporting this idea, we find that performance evaluations are affected by organizing the measures into the balanced scorecard categories when multiple below-target (or above-target) measures are contained within a category but that evaluations are not affected when the above/below-target measures are distributed across the scorecard's four categories.

1. Introduction

In the early 1990s, Robert Kaplan and David Norton (1992) developed a management and measurement tool called the Balanced Scorecard (BSC). The BSC lists a diverse set of performance measures grouped in four categories: financial performance, customer relations, internal business processes, and learning and growth activities (Kaplan & Norton, 1992). Kaplan and Norton (1996a) encourage the inclusion of 4–7 measures in each category. Thus, firms adopting the BSC usually increase the number of performance measures they use and identify a much broader group of measures than those they have traditionally used.

The stated purpose in developing a managerial tool that includes a large number and broad group of performance measures is to improve managerial decision making. While determining whether the BSC improves managers' judgments and decisions can be difficult, a reasonable starting point is to determine whether and how the BSC *affects* these judgments. Prior judgment and decision making research provides evidence of human information processing limitations and decision strategies. We describe and test how these will affect use of the BSC and resulting judgments.

Research in cognitive psychology shows that people are generally unable to process more than 7–9 items of information simultaneously (Baddeley, 1994; Miller, 1956). The BSC contains many more measures than this limit, suggesting that managers will find it difficult to utilize the information in the scorecard. However, the four category organization of the BSC may assist managers' use of this large volume of measures by suggesting a way to combine and use the data. Specifically, decision makers may use a 'divide and conquer' strategy (Shanteau, 1988) where measures within each

* Corresponding author. Tel.: +1-405-325-2293; fax: +1-405-325-7348.

E-mail addresses: mlipe@ou.edu (M.G. Lipe), sesalterio@uwaterloo.ca (S. Salterio).

0361-3682/01/$ - see front matter
PII: S0361-3682(01)00059-9

category are used to make an assessment of the category and these four assessments are then combined. In assessing each category, decision makers are primed to see relations among the measures within each group (Hopkins, 1996). When performance on measures within a group is consistent (e.g. consistently above-target), the decision maker may perceive that the measures are related (i.e. not independent) and consequently, reduce the impact of the individual measures on his or her judgment. In contrast, when the same measures are presented without the organizing BSC categories (or are scattered across BSC categories), the perception of relations among these measures and the resulting reduction in decision weights are less likely.

Our results show that when multiple measures within a BSC category show consistent performance (e.g. above-target), managers' evaluation judgments are reliably different from evaluations made using these same measures without the BSC format. These judgment differences disappear when the measures indicating strong performance are distributed throughout the four BSC categories instead of being found in a single BSC category. Although it is difficult to state with certainty that the BSC results in judgment *improvements*, this study provides evidence that the BSC has predictable and understandable effects on judgment. While these grouping effects may occur with other types of categorizations, other groupings have not received the same kind of attention as those in the BSC.

The remainder of the paper is organized as follows. In the next section we will briefly describe the BSC, review applicable judgment and decision making research, and present a two-part research prediction. Section three describes the experimental work used to test our research predictions and the final section summarizes the conclusions that can be drawn from the study.

2. Background

2.1. The balanced scorecard

In a best-selling book Kaplan and Norton (1996a) describe the methods and procedures necessary for implementing a BSC. The BSC, according to Kaplan and Norton, should contain measures related to financial performance (e.g. return on assets), customer relations (e.g. customer satisfaction surveys), internal business processes (e.g. process efficiency measures), and learning and growth in the organization (e.g. employee capability measures). Kaplan and Norton (1993, 1996b) view the scorecard as a strategic management tool that should explicate the drivers of performance, as well as provide measures of performance. This study focuses on the scorecard's use in evaluation and decision making.

2.2. Cognitive limitations and the divide and conquer decision strategy

The balanced scorecard with its large number of performance measures presents a complex task to a manager asked to use the scorecard to evaluate a division's performance. The manager could, theoretically, weight and combine the many measures into an overall evaluation of the business unit but this is, cognitively, a very difficult thing to do. Research in cognitive psychology has repeatedly shown that humans are able to retain and use only a small number of items in working memory (Baddeley, 1994; Miller, 1956). With this limit on working memory, holding 20 or more individual measures in one's head and mentally manipulating them simultaneously is extremely difficult, if not impossible. Thus, the volume of data in a balanced scorecard suggests that it may overload human decision makers with information.

The balanced scorecard's four categories suggest a way for managers to mentally organize the large number of performance measures that may mitigate this cognitive difficulty. Prior studies show that information processing and judgments are affected by information organization (Bettman & Kakkar, 1977; Payne, Bettman, & Johnson, 1993) and by the hierarchies or relations among information items contained in a decision task (Kleinmuntz & Schkade, 1993). For example, Hopkins (1996) showed that placing an item (e.g. preferred stock) in a particular category (e.g. liabilities) caused experienced professionals to perceive that the item was related to others in the category.

These studies suggest that when data items are grouped in ways meaningful to the decision maker, they may be combined prior to further use (Chase & Simon, 1973). Shanteau (1988) describes this method of using information as 'divide and conquer.' The information is divided into groups, an assessment can be made of each group, and these assessments can then be combined. The organization of the BSC lends itself quite naturally to this kind of mental approach.

2.3. Perceived relations among measures

When using the BSC, the initial stage of the divide and conquer decision strategy is to use measures within a category to assess performance in that area (e.g. financial performance). Since the measures have been grouped together, the decision maker will be expecting and seeking relations between them (Hopkins, 1996; Maines & McDaniel, 2000). If performance on these measures confirms this expectation (e.g. by indicating *consistently* good performance), the decision maker may reasonably reduce the decision weight placed on each individual measure due to perceived correlations (nonindependence) of the measures (Banker & Datar, 1989; Feltham & Xie, 1994). In contrast, if measures indicating good performance are scattered across BSC categories (or contained in uncategorized lists of measures), the decision maker is less likely to expect and perceive these measures to be correlated and to make consequent reductions to their decision weights. This is consistent with findings in psychology that people find it difficult to recognize that correlations exist unless they have theories suggesting such relations (Jennings et al., 1982) and with Maines' (1990) findings in accounting that judgmental discounting for information redundancy (i.e. correlation) does not occur unless the judge is alerted to the presence of such relations (see, especially, her experiment three).

This suggests that judgments made with the BSC will differ from those made with uncategorized lists of measures in particular situations: those cases where performance on measures within a category are consistent (i.e. consistently above-target or consistently below-target). Additionally, the above discussion suggests that judgments made with the BSC will not differ from those made with uncategorized lists of measures in situations where multiple above-target (or below-target) measures are scattered across the BSC categories.

Although performance results on the twenty or more performance measures in a BSC may take on any number of patterns, we will test the impact of only the two extreme patterns described above. That is, we will consider a situation where multiple above-target (or below-target) measures are contained within one BSC category and then we will contrast that with the situation where the above-target (below-target) measures are distributed across categories. For these two situations, we will compare the judgments for decision makers with the BSC to those of decision makers using the same measures without the BSC categories. Our research predictions are:

> Evaluations using the balanced scorecard will differ from evaluations based on the same measures without the scorecard organization, depending on the pattern of performance across categories. Specifically:
>
> 1. judgments are likely to be moderated when multiple above-target (or below-target) measures are contained in a single BSC category but,
> 2. judgments are unlikely to be affected when multiple above-target (or below-target) measures are distributed throughout the BSC categories.

The next section describes the experiments and the test results.

3. Method and results

3.1. Overview of experiments

Participants are presented with a case where they are asked to take the role of a senior executive of WCS Incorporated, a firm specializing in retailing women's apparel. WCS has multiple divisions, the two largest of which are the focus of the case

materials. The case introduces the managers of the two business units and the strategies of the units are described. Multiple performance measures are presented in patterns and formats depending on the experimental treatment as described below. The participant is then asked to evaluate the performance of each of the two unit managers on a scale with seven descriptive labels and numerical endpoints of "0" and "100" (see Table 1 for a sample evaluation form).

After providing the manager evaluations, the participants complete a questionnaire. This questionnaire asks for demographic information, provides manipulation checks (discussed further in Sections 3.2.3 and 3.3.2), and gathers data regarding task difficulty, realism, and understandability.

In both experiments the two divisions described are RadWear and PlusWear, retail divisions specializing in clothing for the urban teenager and in large-sized clothing, respectively. The participants are informed that management believes the performance measures for each division are appropriate for retailers and capture the two different strategies.

3.2. Experiment one

In experiment one we focus on whether the BSC format makes a difference in divisional manager performance evaluation when particularly good or bad performance is contained in one BSC category. In this situation, we predict that the BSC categorization primes the evaluator to perceive consistent performance as evidence of correlation among measures, which may reduce the impact of the individual good or bad measures. This perceived correlation will moderate judgments relative to those made without the BSC organization.

3.2.1. Subjects, design, and procedures

Seventy-eight MBA students served as experimental participants. The students had, on average, 4 years of work experience and 62% were male.

All participants received a diverse set of performance measures, a description of how the

Table 1
Sample evaluation form employed in both experiments

WCS Inc.
Initial Evaluation Form

Year: 1996
Manager: Chris Peters
Division: RadWear
Evaluator: _____

1. Indicate your initial performance evaluation for this manager by placing an 'X' somewhere on the scale below. Note that some label interpretations are provided below.

```
 0                        50                        100
 |----|----|----|----|----|----|----|----|----|----|
Reas-   very   poor     average     good  very   excellent
sign    poor                              good
```

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 2
RadWear balanced scorecard[a] (PlusWear items in parentheses)

Measure	Target	Actual
Financial		
1. Return on sales	24% (22)	25% (23)
2. Sales growth	35% (30)	38% (33)
3. New store sales (new lines sales)	30% (25)	26% (22)
4. Market share relative to retail space	$80 (70)	$80 (70)
5. Return on expenses	42% (36)	42% (36)
Customer-related		
1. **Repeat sales**	30% (40)	33% (36)
2. **Customer satisfaction rating**	95 (97)	96 (96)
3. **Mystery shopper program rating**	96 (96)	98 (94)
4. **Returns by customers as % of sales**	10% (7)	9% (8)
5. Out of stock items	10% (14)	10% (14)
Internal business processes		
1. Average major brand names/store (average % of product range)	32 (88%)	34 (90%)
2. Sales from new market leaders (sales from top brand names)	25% (28)	22% (25)
3. Returns to suppliers	5% (3)	5% (3)
4. Average markdowns	15% (12)	15% (12)
5. Voided sales transactions	3 (2)	3 (2)
Learning and growth		
1. Hours of employee training/employee	10 (8)	11 (9)
2. Average tenure of sales personnel	1.4 (2.1)	1.2(1.9)
3. Employee suggestions/employee	2 (2)	2 (2)
4. Sales personnel taking manager test	30% (36)	30% (36)
5. Stores computerizing	85% (85)	85% (85)

[a] DIFFerent measures are indicated here in bold.

measures were calculated, and the comparison of each measure to its expectation or target for each of the two divisions (see Table 2 for the BSC version of the task).[1] Further, all participants were told that the performance measures were "carefully chosen to represent important aspects of a business unit['s performance]" and were "drivers of the unit's success and linked to its strategy and mission."

The between-subjects (Ss) manipulation was the organization of the performance measures. The BSC group received the 20 measures divided into the four BSC categories (financial measures, customer satisfaction measures, operational measures, and learning measures) while other participants received the same set of 20 measures without the BSC format (NOFORM group). For the NOFORM group the measures were presented in one of two orders, alphabetical or random.[2] In addition to the format manipulation across subject groups, the order of presentation of the two divisions (i.e. RadWear and PlusWear) was counterbalanced across subjects within each format group.

For all participants, the financial measures indicated that performance was somewhat above expectations for both divisions (note in Table 2 that two financial measures were above-targets,

[1] Participants received separate exhibits for RadWear and PlusWear (and none of their measures were shown in bold). Measures for both divisions are included in Table 2 for efficiency of exposition.

[2] The order of measures for the latter was chosen by random draw with the only proviso that adjacent measures should not come from the same BSC category. Two orders were used for the NOFORM group to increase the generalizability of results.

one below-target, and two on-target). Further, for all participants, one division was *above* expectations in its customer related measures and the second division was *below* expectations in the customer-related measures (note that Table 2 shows four RadWear customer measures better than target and four PlusWear worse than target; these items are shown in bold in the table). The two remaining groups of measures (internal business processes and learning and growth) were approximately at expectations for all participants (note that Table 2 shows one measure above-target, one below-target, and three on-target). Therefore, there was one within-subjects manipulation: the division's being above (positive performance) or below (negative performance) the customer-related performance measures targets.[3]

As noted above, performance relative to target was similar across the two divisions for all performance measures except for four customer-related measures (shown in bold in Table 2). We will refer to these as the DIFFerent measures. In the BSC format, these four measures were grouped together in the second category. Thus, in the BSC format the two divisions were performing equally on three of the four dimensions, with RadWear superior on the other. In contrast, for the NOFORM group, the 20 measures were not grouped into categories. Instead the measures were listed in an alphabetical or random order,[4] neither of which suggests that particular measures are correlated. These two NOFORM orders resulted in the DIFFerent measures being in positions 4, 8, 11, and 12 (out of 20) for the alphabetical listing and in positions 2, 4, 12, and 19 for the random listing.

Table 3
ANOVA results for experiment one manager evaluations

Variable	df	SS	MS	*F*	*P*
Between Ss					
Organization	1	41.25	41.25	0.14	0.71
Order	1	4.10	4.10	0.01	0.91
Organ.×Order	1	0.52	0.52	0.00	0.97
Error	74	22,567.86	304.97		
Within Ss					
Division	1	13,917.31	13,917.31	97.14	0.00
Div.×Organization	1	817.27	817.27	5.70	0.02
Div.×Order	1	1513.64	1513.64	10.57	0.00
Div.×Organ.×Order	1	344.02	344.02	2.40	0.13
Error	74	10,601.94	143.27		

3.2.2. Dependent measure

All subjects evaluated each manager using the evaluation form and scale shown in Table 1. We expect that there will be a main effect for division, showing that differential divisional performance on the customer-related measures affects their managers' evaluations. Additionally, we expect an interaction of organization and division, showing that the BSC organization moderates the evaluations of the two divisional managers relative to evaluations without the BSC organization, given that multiple below-target (for PlusWear) or above-target (for RadWear) measures are contained in one BSC category (i.e. customer-related measures).[5]

3.2.3. Results

Checks on the effectiveness of the manipulations revealed that participants receiving the BSC format felt that the performance measures were more logically organized and usefully categorized than those receiving the NOFORM performance measures (both P-values < 0.01). No other differences were noted for these groups for questions regarding difficulty of the task, emphasis on financial measures, or extensiveness of measures provided (all P-values > 0.10). Within the NOFORM group, no differences were found for subjects with the

[3] While academic research has produced mixed results regarding the impact of customer satisfaction on profitability (e.g. Foster & Gupta, 2000; Ittner & Larcker, 1998), managers generally believe that customer satisfaction is a key performance driver, especially in the retail sector (Rucci, Kirn, & Quinn, 1998). In our experiment, participants were told that all measures chosen for the BSC were drivers of the unit's success.

[4] It should be noted, however, that in either case, after each five measures, a blank line was inserted in the list so that readability and eye fatigue would not differ for the NOFORM and BSC formats.

[5] Since judgments are strongly affected by comparison cases (Hsee, 1996, 1998), we expect that information organization will most likely affect the comparative or *relative* judgments regarding the two managers.

Table 4
Descriptive statistics for experiment one manager evaluations—means (standard deviations)

Scorecard format	RadWear	PlusWear	Evaluative difference (RadWear−PlusWear)
BSC	69.77 (14.21)	52.50 (16.93)	17.27 (20.32)
NOFORM	74.32 (10.24)	50.03 (18.02)	24.29 (15.03)
Both formats	71.76	51.42	

alphabetic versus the random order for any of these questions (all P-values > 0.10) or for the managerial evaluations. Also, the order of the presentation of divisions had no effects on responses to the manipulation check questions (all P-values > 0.10). Although division order was not related to the hypotheses, it did interact with division in affecting performance evaluations ($F = 10.57$, $P < 0.01$). Thus, division order is included in the statistical analysis but it is not discussed further.[6]

Analyzing the individual manager evaluations via a repeated measures $2 \times 2 \times 2$ analysis of variance (ANOVA) with scorecard organization and division order as between-Ss factors and division as a within-Ss factor (see Table 3), indicates statistically significant effects for division ($F = 97.14$, $P < 0.01$) and the interaction of division and organization ($F = 5.70$, $P < 0.02$). These results show that RadWear's manager is evaluated higher than PlusWear's (means [standard deviations] of 71.76 [12.76] and 51.42 [17.34], respectively) and that the scorecard's organization affects the relative evaluations of the two divisional managers.

The interaction of organization and division supports our first research prediction, showing that information's organization affects the relative evaluations of the managers for this pattern of performance results where particularly positive/negative performance is concentrated in one BSC category. As shown in Table 4, participants with the four category organization of measures evaluated RadWear's manager 17.27 points higher than PlusWear's while participants with the NOFORM format evaluated RadWear's manager 24.29 points higher than PlusWear's. When the multiple positive/negative measures all related to customer relations, they led to less extreme evaluations when they were displayed in this BSC category as opposed to being distributed throughout the unorganized list.

3.2.4. Supplemental analysis

Since our research predictions were based on the idea that subjects with BSC categories would use a divide and conquer strategy, we expect to see differences in the patterns of data processing for BSC versus NOFORM subjects. To test this, 64 of the subjects in experiment one provided memos explaining each of their managerial evaluations. The NOFORM subjects mentioned, on average, 22.6 individual performance measures in these memos. They also referred to self-generated groups of measures 1.1 times, on average, with the most common group of measures mentioned being a self-generated customer-related grouping. In total, about 95% of the items mentioned by these subjects were individual measures. The BSC subjects referred to an average of 18.7 individual measures. They also mentioned 8.1 groups of measures (including multiple mentions of some groups) and most of these (98%) were the BSC categories. The most common group mentioned was the customer-related BSC category. The patterns of usage of group versus individual measures are significantly different for the BSC and NOFORM subjects [$\chi^2(1) = 249.16$, $P < 0.01$]. Logically, the BSC organization led to increased consideration of groups of measures. Thinking about the measures in these groups led to moderated judgments when measures with above-target (or below-target) performance were contained within a single BSC category and may have been correlated (nonindependent).

[6] Specifically, when PlusWear was rated first, the average evaluations were 67.97 and 54.97 for RadWear and PlusWear, respectively. When RadWear was rated first these average evaluations were 74.39 and 48.96.

3.3. Experiment two

In experiment two we focus on whether the BSC format makes a difference in divisional manager performance evaluation when particularly positive (or negative) performance measures are distributed across all four BSC categories. Experiment one shows that the BSC organization can impact judgments given a particular pattern of performance results (i.e. where consistently positive or negative performance is found in one category). In contrast, our second research prediction posits that the BSC's organization will not affect judgments when the multiple positive/negative measures are distributed across categories (so that the categorization does not suggest the DIFFerent measures are related).

3.3.1. Subjects, design, and procedures

Seventy-one students in graduate-level managerial accounting courses participated as subjects. The students had average work experience of 5 years and 58% were male. The dependent measure for experiment two is the same as that of experiment one, the managerial evaluations.

Similarly to experiment one, a 2×2×2 repeated measures design was used. Performance measure organization was varied between Ss; students either saw the measures in the BSC format or the NOFORM (random only) format. The order of presentation of RadWear and PlusWear was again varied between Ss. The within-Ss factor was divisional performance on four DIFF measures, with RadWear superior to PlusWear on all four. In contrast to experiment one, here these DIFF measures were distributed across the BSC categories. They were Return on Sales, Mystery Shopper program rating, Average major brand names/store (or Average% of product range), and Hours of employee training/employee. For all other measures, RadWear and PlusWear performed similarly relative to their targets; for the financial measures, both divisions were above target on one (non-DIFF) measure, below target on one, and met the target on the other two. This pattern was repeated for the customer-related category and for the internal business processes group. In the learning and growth category, both divisions beat their targets for one (non-DIFF) measure and met their targets on the other three.[7] The DIFF measures were in positions 1, 7, 13, and 19 (out of 20) in the BSC format and 2, 4, 12, and 19 in the NOFORM format. This NOFORM positioning is the same as those used in the random listing in experiment one.

3.3.2. Results

As in experiment one, subjects with the BSC format judged the performance measures to be more logically organized and usefully categorized than those with the NOFORM format (both $P<0.01$). The 2×2×2 repeated measures ANOVA indicates only two statistically significant effects: Division ($F=24.30$, df = 1.67, $P<0.01$) with RadWear evaluated higher than PlusWear (means [standard deviations] of 69.49 [11.97] and 61.44 [14.36], respectively) and Order ($F=3.91$, df = 1,67, $P=0.05$).[8] The format of the performance measures did not affect the judgments ($F=0.20$), nor did it interact with division ($F=0.48$). Thus, in contrast to results for experiment one, with this pattern of performance results (i.e. with the above/below-target measures distributed across BSC categories), the BSC format did not affect the evaluations of the managers.

The experiments indicate that, dependent on the pattern of performance results, organizing measures into the BSC can affect managerial judgments. This may be caused by information

[7] Although it would have been ideal to use the same DIFF measures in experiments one and two, it was not possible to do this while credibly placing these measures all into one BSC category in one while distributing them across categories in the other. Instead, experimental control was maintained by holding many other things constant across the two experiments. For example, in both experiments RadWear (PlusWear) beat its target on eight (four) measures, missed it on three (seven), and tied it on nine (nine). Summing up the percentage above or below target on each measure, RadWear (PlusWear) beat its targets by a sum of 12.5% (−30.31%) in experiment one and 13.46% (−27.05%) in experiment two. The difference in summed percentage from target for RadWear minus PlusWear was 42.81% in experiment one and 40.51% in experiment two. Thus, the relative performance of the two divisions was similar across the experiments.

[8] When PlusWear was evaluated first, mean evaluations were 68.09 versus 62.92 when RadWear was evaluated first. Order did not interact with any other factors.

processing strategies based on the BSC categorization highlighting the potential relations among measures within categories.

4. Limitations and conclusion

Our experimental design has several limitations. First, many of our experimental participants were novices in the use of the BSC and they did not necessarily have business experience in the retail sector from which we pulled much of our case materials. Although the information processing effects we tested relate to basic issues of cognition, we do not know whether or how further experience may impact the observed effects.[9] For example, persons with more knowledge of the retail apparel industry may have a better sense of the sensitivity of each measure to managerial actions (perhaps affecting weightings on the measures) and to correlations among measures throughout the scorecard (again, affecting weightings). Thus, the BSC categories may have less effect on these experts, who do not need the categorical organization to perceive and use these relations. Of course, even experts will have a learning period, during which they may act much like our experimental participants. Nonetheless, our experimental results provide a baseline against which a study of BSC experts could be compared.

A second limitation of our study is that we cannot assess the accuracy of our participants' evaluations since there is no accepted normative model for determining performance evaluation scores (see also Lipe & Salterio, 2000). The direction of the effect, however, seems consistent with the normative response to nonindependence of measures (e.g. Feltham & Xie, 1994).

A third limitation of our study is that we only investigate subjective performance evaluations (Kaplan & Norton, 1996a, pp. 217–223). In some companies explicit weightings or formulas are used for combining performance measures to determine evaluations (e.g. Malina & Selto, 2000) and in some cases an explicit two stage process is used. In the first stage an evaluation is made on a category-by-category basis and in the second stage the manager combines these category level judgments into an overall performance evaluation (e.g. Ittner, Larcker, & Meyer, 1998). Both approaches are examples of managerial decision aids (i.e. explicit weights and formulas) to deal with information processing limitations.

While we focus on the classification scheme provided by Kaplan and Norton's (1996a) BSC, other categorizations of performance measures may lead to similar results. That is, as long as the categories have meaning to the decision maker they may prime him or her to seek relations among measures and to react to any perceived correlations by reducing the impact of individual measures. Further research should explore whether the BSC categorization has any unique effects, relative to other meaningful categorization schemes. Again, our study provides a baseline against which future studies can be compared.

The balanced scorecard has received significant attention in the business press and a recent survey estimates 60% of Fortune 1000 firms have experimented with the BSC (Silk, 1998). Despite this widespread attention and use, the interaction of the BSC and managers' cognition has received little consideration. Many judgment issues deserve research attention. For example, how are trade-offs made across BSC categories, how are the individual measures in a category weighted, and how does the differential reliability, precision, and other characteristics of the measures affect the weight placed on them? Other important empirical questions include what is the covariation structure among the BSC measures within and between categories, how are the categories linked to underlying dimensions of managerial performance (i.e. how sensitive are they to managerial actions), and is it common for one distinct aspect of managerial performance to affect measures in multiple BSC categories? Answers to these questions will in turn lead to further research questions regarding the judgmental effects of the BSC.

[9] Twenty of the subjects in experiment two indicated they had experience working for a BSC firm. Including experience as a covariate in the analysis had no effect on the test results.

Acknowledgements

We are grateful to the Canadian Academic Accounting Association for project funding. Helpful comments were provided by Joan Luft and participants of the New England Behavioral Accounting Research Consortium, American Accounting Association Annual meeting, Canadian Academic Accounting Association Annual Conference, and the Universities of Alabama, Arkansas, Brigham Young, and Wilfrid Laurier. We thank Karen Cravens and Peter Tiessen for help in recruiting subjects.

References

Baddeley, A. (1994). The magical number seven: still magic after all these years? *Psychological Review*, *April*, 353–356.

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

Bettman, J. R., & Kakkar, P. (1977). Effects of information presentation format on consumer information acquisition strategies. *Journal of Consumer Research*, *March*, 233–240.

Chase, W. G., & Simon, H. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing*. New York: Academic Press, pp. 215–281.

Feltham, G., & Xie, J. (1994). Performance measure congruity and diversity in multi-task principal/agent relations. *The Accounting Review*, *69*(3), 429–454.

Foster, G., & Gupta, M. (2000). *The customer profitability implications of customer satisfaction*. Stanford working paper.

Hopkins, P. (1996). The effect of financial statement classification of hybrid financial instruments on financial analysts' stock price judgments. *Journal of Accounting Research, supplement*, *34*, 33–50.

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.

Hsee, C. (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.

Ittner, C., & Larcker, D. (1998). Are nonfinancial measures leading indicators of financial performance? An analysis of customer satisfaction. *Journal of Accounting Research*, *36 supplement*, 1–46.

Ittner, C., Larcker, D., & Meyer, M. (1998). *The use of subjectivity in multi-criteria reward systems*. Wharton School working paper.

Jennings, D., Amabile, T., & Ross, L. (1982). Informal covariation assessment: Data-based versus theory-based judgments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: heuristics and biases* (pp. 211–230). Cambridge: Cambridge University Press.

Kaplan, R., & Norton, D. (1992). The balanced scorecard—measures that drive performance. *Harvard Business Review*, *January–February*, 71–79.

Kaplan, R., & Norton, D. (1993). Putting the balanced scorecard to work. *Harvard Business Review*, *September–October*, 134–147.

Kaplan, R., & Norton, D. (1996a). *The balanced scorecard*. Boston, MA: Harvard Business School Press.

Kaplan, R., & Norton, D. (1996b). Using the balanced scorecard as a strategic management system. *Harvard Business Review*, *January–February*, 75–85.

Kleinmuntz, D., & Schkade, D. (1993). Information displays and decision processes. *Psychological Science*, *4*(4), 221–227.

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

Maines, L. (1990). The effect of forecast redundancy on judgments of a consensus forecast's expected accuracy. *Journal of Accounting Research*, *28*(supplement), 29–47.

Maines, L., & McDaniel, L. (2000). Effects of comprehensive-income characteristics on nonprofessional investors' judgments: the role of financial-statement presentation format. *The Accounting Review*, *75*(2), 179–207.

Malina, M. A., & Selto, F. H. (2000). *Communicating and controlling strategy: an empirical study of the effectiveness of the balanced scorecard*. University of Colorado at Boulder working paper.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for information processing. *The Psychological Review*, *March*, 81–96.

Payne, J., Bettman, J., & Johnson, E. (1993). *The adaptive decision maker*. Cambridge: Cambridge University Press.

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

Shanteau, J. (1988). Psychological characteristics and strategies of expert decision makers. *Acta Psychologica*, *68*, 203–215.

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