

**COMMON-MEASURES BIAS IN THE BALANCED SCORECARD: COGNITIVE
EFFORT AND GENERAL PROBLEM-SOLVING ABILITY**

Aleecia R. Hibbets*
The University of Louisiana Monroe

Michael L. Roberts
University of Colorado at Denver

Thomas L. Albright
The University of Alabama

*Corresponding Author:

700 University Avenue
Monroe LA 71209
318-342-1115
hibbets@ulm.edu

November 1, 2006

COMMON-MEASURES BIAS IN THE BALANCED SCORECARD: COGNITIVE EFFORT AND GENERAL PROBLEM-SOLVING ABILITY

ABSTRACT

Lipe and Salterio (2000) found decision-makers role-playing as superiors in a performance evaluation setting placed 97 percent of criteria weight on information common to subordinates in making their judgments when evaluating performance using the Balanced Scorecard (BSC). Because theory underlying the BSC (Kaplan and Norton 1996, 2001) assumes all measures included are relevant for achieving organizational objectives, several subsequent studies have attempted to understand or alleviate this common-measures bias (e.g., Krumwiede et al. 2002; Banker et al. 2004; Libby et al. 2004; Roberts et al. 2004; Dilla and Steinbart 2005). This study is the first to directly test the effort-based explanation of the common-measures bias (Slovic and MacPhillamy 1974, Lipe and Salterio 2000). Additionally, this research investigates the role of participants' problem-solving ability on mitigating the common-measures bias. Results show strong support for the effect of problem-solving ability on increased measures usage, but not for the effort-based explanation. Superiors who relied more on common measures were less convinced subordinates should be evaluated with different measures and less convinced that subordinates' target markets were unique. These results raise important questions about the nature of common-measures bias.

I. INTRODUCTION

Kaplan and Norton's Balanced Scorecard (BSC) has been promoted as an integrated, balanced approach to performance measurement and improvement in which multiple organizational goals are measured and managed simultaneously to produce desired results (Kaplan and Norton 1996). At its core is a shift of focus from the use of financial results as the sole source of performance measurement and reporting, to balancing financial targets with three categories of nonfinancial drivers of success: customer-related, internal business processes, and learning and growth. However, Lipe and Salterio (2000) detailed disproportionate weighting of common measures by BSC users in a performance evaluation context. Their experimental participants relied almost exclusively (97 percent weight) on the common measures when evaluating subordinate managers. A common-measures effect is especially problematic in a BSC context because the BSC is designed to be tailored to the unique strategy of the organizational unit, and both common and unique performance measures are important drivers of the success of the organization.

Common-measures bias has most often been explained as decision-makers' unwillingness to incorporate the unique information because this information requires greater cognitive effort to process (see, for example, Slovic and MacPhillamy 1974, Lipe and Salterio 2000). This study tests that explanation by requiring additional cognitive processing. Prior theory suggests evaluating performance on each measure prior to making an overall evaluation, one of the stages Slovic and MacPhillamy theorized as an increased cognitive load on the decision-maker, should increase the weight placed on unique information. That is, if the relative performance on unique measures has already been processed by decision-makers in an earlier phase of the task, it should be cognitively easier to incorporate in the overall performance

judgment phase. Thus, one-half of the participants were instructed to evaluate the performance of each division manager on each BSC item individually prior to making overall performance evaluations of the two division managers.

In addition, we examine the potential contribution of general problem-solving ability on unique vs. common measures usage. Kennedy (1993) theorized that decision biases could be reduced by replacing the decision maker with someone possessing greater mental capacity for processing. Thus, we test whether BSC users who exhibit greater general problem-solving ability are able to utilize more of the information contained in a BSC performance report.

The results support problem-solving ability as an explanation for more versus less information usage. Evaluators with greater general problem solving ability utilized more of the unique measures than evaluators with lesser general problem solving ability. However, the cognitive effort explanation proposed by Slovic and MacPhillamy (1974) for overweighting common measures was not supported.

Supplemental analysis also revealed participants who relied more heavily on common measures significantly differed from other participants on two important follow-up questions. The former were (1) less certain the two divisions were targeting different markets and (2) were also less convinced the two divisions should use different performance measures. These are interesting results since they raise questions about the basic nature of the common-measures bias. Specifically, they suggest BSC users may intentionally and consciously choose to rely on the subset of common measures as a rational decision strategy rather than the effect being a judgment bias in which decision-makers perhaps unknowingly emphasize the common measures (see, for example, Slovic and MacPhillamy 1974). Unfortunately, the current study is merely able to raise this possibility; future research is needed to further investigate this possibility.

The next section reviews relevant research related to the Slovic and MacPhillamy (1974) and Lipe and Salterio (2000) studies and subsequent research and develops the research hypotheses. Then the methodology is described. Following this, results of hypotheses are reported and the findings are discussed. The paper concludes with limitations and opportunities for future research.

II. LITERATURE REVIEW AND HYPOTHESES DEVELOPMENT

In a series of five experiments, participants in Slovic and MacPhillamy (1974) compared pairs of students on one common and one unique dimension. Across all five experiments, slightly more weight was given to the common factor than to the unique factor. Neither cautioning the judges to avoid relying on the common dimension nor providing subjects with correct answer feedback reduced the effect.

Although the experiments clearly demonstrate the common-measure effect on decisions, the underlying cause as well as the “basic mechanisms” (1974, 192) of this effect are unclear. As a possible explanation, the concept of cognitive difficulty or strain is primarily advanced; that is, the unique information is more difficult to process and incorporate into a judgment, and is thus discounted by the judge “in an attempt to reduce the strain on memory, attention, and other components of reasoning” (1974, 173). The authors state comparing the students on the common dimension “should be easier, cognitively, than a comparison between dimensions. This ease of use could lead to greater reliance on the common dimension” (1974, 174)¹. Research in

¹ Two other possibilities were suggested to explain the observed common-measures effect. First, subjects may “discount the unique information because they lack confidence in their ability to use it appropriately. In other words, the ease of comparison on the common dimension could induce greater confidence in one’s ability to use that information, and this confidence, in turn, could mediate the weight given the information” (192). Secondly, participants may substitute the average value for the missing item for each student, resulting in a smaller difference between stimuli on the unique dimensions, which then is discounted by the decision maker. A bias in favor of the common dimension would occur. The authors report a “number of subjects” volunteered comments indicating “that

organizational and consumer behavior also has demonstrated a bias in comparative evaluations towards common measures (Markman and Medin 1995; Zhang and Markman 1998; Kivetz and Simonson 2000).

This phenomenon was also demonstrated in a performance evaluation setting using the BSC. Lipe and Salterio (2000) developed an experimental case based on Kaplan and Norton's (1996) Kenyon Stores example of a retail firm utilizing the BSC. Two divisions, RadWear and WorkWear, each target specific markets and have unique strategic objectives. The case described the development of a BSC consisting of sixteen performance metrics for each division through the collaboration of the division manager and top management.

The experimental participants in Lipe and Salterio (2000) were presented with the division scorecards and asked to assume the role of a senior manager to evaluate the performance of the two division managers. Results showed the evaluations were systematically influenced by the pattern of performance on the common measures, but not affected by the pattern of performance on the unique measures. The division with superior performance on the common measures was rated an average of six to seven points higher than the other division. However, there was no significant difference in the evaluations of the managers who performed better on the unique measures. In a repeated-measures ANOVA, only the DIVISION x COM [common] interaction term was significant ($p < 0.01$), indicating the pattern of performance on common measures affected the evaluation, while the pattern of performance on unique measures did not. In addition, regression analysis using the difference in manager ratings as the dependent variable revealed a significant, positive beta of 10.87 ($p < 0.01$) for common information, but an insignificant slope coefficient of 0.08 ($p > 0.10$) for unique information.

they did tend to assume average values for missing scores and did discount small differences between stimuli on a dimension" (192).

Since Lipe and Salterio (2000), several studies have demonstrated BSC users can incorporate unique information into performance judgments to a significant degree if participants (1) first disaggregate the BSC into separate judgments (Roberts et al. 2004), (2) are experienced business professionals explicitly reminded about the importance of all of the BSC measures (Roberts et al. 2002), (3) are trained in the intent and design of the BSC (Dilla and Steinbart 2005), (4) view multiple periods of performance data and receive outcome feedback (Krumwiede et al. 2002), (5) receive detailed strategy information to highlight the direct linkage of the unique measures to achieving divisional objectives (Banker et al. 2004), or (6) are required to justify their performance evaluations or receive explicit assurance regarding the reliability of the data (Libby et al. 2004). In the current study, these prior findings provide a foundation for investigating the role of increased effort and problem-solving ability in participants' usage of unique measures.

Testing an effort-related explanation

If Slovic and MacPhillamy's (1974) rationale of increased cognitive effort is correct, then common measures bias should be reduced or eliminated if decision-makers expend additional effort to evaluate performance on each scorecard item, both common and unique, prior to making a global performance judgment. This possibility is consistent with a cognitive/attentional resource allocation perspective (Kanfer and Ackerman 1989; Kanfer et al. 1994), which views task performance as jointly affected by an individual's attentional resource capacity, the attentional resource demands of the task, and the individual's allocation of attentional resources across both on- and off-task activities. According to this theory, effort is the amount of attentional resources devoted to a task (Norman and Bobrow 1975).

Kanfer and Ackerman (1989) described two processes a decision-maker uses in allocating cognitive resources: distal and proximal processes. Distal processes initially occur prior to engaging in the task, and involve “the choice to engage any, some, or all of one’s resources for attainment of a goal” (1989, 661). Three perceptions of functional relationships determine distal decisions: the perceived (a) performance-utility, (b) effort-utility, and (c) performance-resource functions. The proximal processes then “determine the distribution of effort across on-task and off-task activities during task engagement” (1989, 662) and include the self-regulatory activities of self-monitoring, self-evaluation, and self-reaction. Underlying these concepts is the assumption that an individual decision-maker has a limited availability of cognitive attentional resources (Kanfer et al. 1994).

If the decision-maker decides to engage cognitive resources to consider each performance measure individually, both common and unique, in an earlier phase of the task, subsequently utilizing these measures, whether common or unique, in the performance evaluation task should be attentionally less expensive, promoting inclusion of all performance measures in the evaluation. The distal processes involved first lead a decision-maker to choose to engage resources in the “individual item assessment” phase of the BSC task, and secondly, induce the choice to again employ the cognitive resources needed to integrate all the individual measures into a performance judgment. This expectation is derived from the decision-maker’s consideration of the perceived effort-utility function, since after the initial phase of the task, the effort required (and thus, perceived costs) is reduced. The reduction in effort is expected to produce renewed commitment to completion of the task without resorting to a simplifying strategy.

Kennedy's (1993, 1995) debiasing framework is also applicable to mitigating the common-measures bias. Broadly, Kennedy distinguished between two sources of bias, effort-related and data-related, though these two classes are not mutually exclusive (1993, 234). It is not clear whether the cognitive simplifying strategy believed to have motivated Lipe and Salterio's (2000) participants was due to an effort-related or data-related bias. An effort-related bias is suggested, though, since (1) participants chose to ignore unique, but not common measures (suggesting the selection of measures to include in the decision was not arbitrary); (2) these results are consistent with Slovic and MacPhillamy's (1974), in which the task was much less demanding than Lipe and Salterio's (2000), and (3) participants did not have to rely on working memory, but could have made notes or calculations during the task.

Evidence from prior research suggests that requiring evaluators to increase effort to understand individual measures can lead to placing more weight on unique BSC measures. Dilla and Steinbart's (2005) undergraduate business students, who were trained in the purpose and design of the BSC, placed greater weight (eta-squared values of 0.078 versus 0.012) on unique measures compared to Lipe and Salterio (2000). This suggests that prior information processing of BSC data encourages decision-makers to forego the simplifying strategy of the common-measures bias. Roberts et al. (2004), which required participants to evaluate performance on each BSC objective separately, assigning each BSC item an individual performance rating before making a global performance evaluation, showed that superiors' overall performance ratings of each division manager reflected both common and unique information consistently with prespecified weights adopted by the firm. Taken together, these results suggest an effort-related common-measures bias, consistent with Lipe and Salterio's (2000) explanation for their results. However, no study has directly investigated whether an effort-based explanation holds for the

common-measures bias. The current study thus extends Roberts et al. (2004) by being the first research to directly test the cognitive effort theory for the common-measures bias by manipulating whether participants are required to analyze performance on individual BSC measures prior to making a global performance judgment.

In accordance with an effort-related bias explanation, this study investigates whether including an evaluation of each individual performance item prior to requesting the manager's overall performance evaluation causes decision-makers to expend the effort necessary to overcome the common-measures bias. If Lipe and Salterio's (2000) participants chose to evaluate only the common measures to simplify the cognitive demands, then our participants, having already evaluated the unique measures in the first phase of the task, are expected to perceive the incremental cost of including the unique measures in the performance evaluation to be sufficiently reduced so that both common and unique information will be reflected in their evaluations of division performance. The first hypothesis, stated in alternative form, is:

H1: Managers who evaluate performance on individual BSC measures before making an overall performance evaluation will place more weight on unique performance measures, i.e., the difference between the two managers' ratings will be smaller, in their overall performance judgments than those who assess individual BSC items after making an overall performance evaluation.

There are several reasons why this hypothesis may not hold. First, participants may choose to underweight some items based on their belief that those items do not reflect the strategy of the division being evaluated (Banker et al. 2004). Second, it is possible that graduate business students lack the experience and expertise to consolidate sixteen common and unique performance measures into a performance judgment, possibly because they are unable or

unwilling to do so (Roberts et al. 2002). Third, participants may discount some measures if they believe they are less controllable by the manager (Krumwiede et al. 2004). Fourth, participants may need more than a one-period setting to fully understand the necessary information contained in the unique measures, since experiencing multiple periods of data provides outcome feedback which can be used for belief revisions (Krumwiede et al. 2002). This study included only one period of performance data.

The role of ability

Cognitive abilities represent one of the basic determinants of learning and performance (Kanfer and Ackerman 1989). Prior research in accounting demonstrates the importance of problem-solving ability on task performance (Bonner and Lewis 1990), especially for relatively unstructured, complex tasks (Bonner et al. 1992; Tan and Kao 1999). The BSC, with a total of sixteen to twenty-eight measures, lack of normative guidelines to evaluate the accuracy of performance evaluations, and lack of a combination rule to assimilate the information cues into a judgment, represents a complex, ill-structured task (Dilla and Steinbart 2005).

Kennedy's (1993, 1995) concept of capacity corresponds with ability; it is included under effort-related biases, and its inclusion in the framework implies potential limitations of a decision-maker. Kennedy does not provide detailed guidance for increasing judgment quality when there are concerns regarding capacity or ability, assuming "that capacity is sufficient or can be augmented if required" (1995, 251). At a minimum, though, including capacity in the framework implies both individual differences in ability and the importance of ability as a determinant of judgment quality. The cognitive resource allocation perspective (e.g., Kanfer et al. 1994) also speaks to ability with the concept of limited attentional resource capacity. Kanfer and Ackerman (1989) reference Ackerman (1984, 1986, 1987) as suggesting that "individual

differences in general intellectual ability may be conceptualized as differences in individuals' total attentional-cognitive [resource] capacity" (1989, 663).

Kanfer and Ackerman (1989) denote three phases or classes of ability important for predicting individual differences in performance during skill acquisition. Of the ability types, general intellectual abilities (including reasoning and broad content abilities such as verbal, numerical, and spatial) are attributed to be most closely associated with the first phase of information processing, acquiring declarative knowledge² (Kanfer and Ackerman, 1989). When a task places substantial demands on the attentional system, correlations between general intellectual abilities and task performance are correspondingly high (see, for example, Ackerman 1988).

Since ability has been shown to be an important determinant of judgment performance, we posit that managers with greater problem-solving ability³ will be capable of synthesizing multiple information cues into a holistic judgment and will be better able to use all provided information (both common and unique measures) in determining a performance evaluation. To our knowledge, problem-solving ability has not been investigated in the area of BSC evaluations and common-measures bias. The second hypothesis, stated in alternative form, is:

H2: Managers with greater problem-solving ability will place more weight on unique performance measures, i.e., the difference between the two managers' ratings will be smaller, in their judgments.

This hypothesis may not hold if problem-solving ability is not the type of ability attributable to performance in this task, and instead general intellectual abilities (Ackerman

² Kanfer and Ackerman (1989) further associate "perceptual speed abilities" with the second phase of skill acquisition, procedural knowledge, and "psychomotor abilities" with the final, "automatic" stage of skill acquisition (664).

³ Rather than investigating general intellectual ability, we follow the work of prior accounting studies (e.g., Bonner and Lewis 1990) and instead focus on the role of problem-solving ability.

1988) or another more specific type of ability for integrating information cues is more applicable for this task.

III. METHODOLOGY

Similar to prior studies (e.g., Lipe and Salterio 2000), graduate business students completed a classroom case describing WCS, Inc., a clothing firm, and two of its retail divisions, RadWear and WorkWear. We used the same four-category BSC developed by Lipe and Salterio (2000). Participants were given an overview of the BSC implementation process at WCS and then presented with each division manager's scorecard containing target and actual performance on sixteen performance measures. Eight measures on each division scorecard were common to the two managers, and eight measures on each scorecard were unique to that manager to align with the division's strategic objectives. Both division managers performed better than target on all scorecard measures, but one division manager outperformed the other on common measures, and one manager outperformed on unique measures. The sum of excess performance (total percentage above target) across all measures, common and unique, was approximately the same for both divisions⁴. To maintain comparability, precautions were taken to ensure the experimental case replicated Lipe and Salterio's (2000) except for the manipulations being tested.

A full factorial design has been replicated by several studies (e.g., Banker et al. 2004; Dilla and Steinbart 2005; Roberts et al. 2004). Focusing on a combination where common-measures bias is possible, i.e., when performance is mixed, is a more efficient use of the pool of participants. Therefore, following Libby et al. (2004), we employed one of the four cases

⁴ Total percent above target for RadWear is 136.38; total percent above target for WorkWear is 137.07, a difference of 0.69. The small difference in total is due to different bases among the performance measures.

developed by Lipe and Salterio (2000): the case where performance on common measures favors RadWear and performance on unique measures favors WorkWear. Role-playing as WCS executives, experimental participants provided evaluations for each division manager.

Subjects

Ninety-eight graduate business students recruited from two universities participated. Thirty-one were MBA students and 67 were Masters of Accountancy students. No systematic differences were observed between the groups of participants. Thus, responses were pooled for analysis. One participant did not provide overall scores for either manager and was excluded from analysis, resulting in 97 useable responses.

Thirty-one participants (32 percent) were male. Mean age was 25.6 years and ranged from 21 to 49 years. Work experience ranged from 0 to 20 years with a mean of 2.8 years. A majority (63 percent) indicated an emphasis in accounting in their graduate studies, and twenty-six participants (27 percent) indicated their full-time work experience was most closely related to accounting, auditing, or taxation. Participants generally rated their knowledge of the BSC methodology as neutral (mean = 0.41 on a scale ranging from -5.0 = “very unfamiliar” to 5.0 = “very familiar”).

Design and Procedure

The experimental design in this study is 2 x 2. The first between-subjects factor, BEFORE, is manipulated at two levels, before and after. One-half of participants were asked to rate each manager’s performance on each scorecard item before making their overall performance evaluation of the two WCS managers; the other half of participants complete the same task after evaluating the division managers’ performance. The purpose of eliciting these individual ratings is to require half of the participants to expend the cognitive effort necessary to

evaluate the relative importance of each measure before completing the performance evaluation task. The explicit ratings participants provided are also available to control for differences in responses and for supplemental analysis.

The second between-subjects factor is general problem-solving ability (GPS). To measure GPS, we relied on one widely-used measure of general problem-solving ability in accounting studies (Bonner et al. 1992). This measure was introduced in Bonner and Lewis (1990) as a twelve-item scale with four questions in each of three subsections (analogical reasoning, data interpretation, and analytical reasoning) taken from the 1987-4 Graduate Record Examination (GRE) to measure general problem-solving ability and its effect on performance. The scale was refined in Bonner et al. (1992), who ultimately reduced the measure to eight items (with a reported coefficient alpha of 0.63) in assessing general problem-solving ability's role in tax issue identification. Nine of the Bonner and Lewis (1990) GRE items were used to create our GPS measure: three in each category of analogical reasoning, data interpretation, and analytical reasoning. Each correct answer was summed for each participant, creating a preliminary variable, GPSSUM. GPSSUM had an observed range of one to nine, with a mean (median) of 4.58 (4.0). To create the variable GPS, GPSSUM was split at the midpoint (4.5), with 48 (49) participants in the high (low) GPS category⁵. The mean number of correct questions for the high (low) GPS group was 6.33 (2.86).

Participants were provided with (1) a brief description of WCS, Inc. and its two largest subsidiaries, RadWear and WorkWear, (2) an overview of the firm's implementation of the BSC,

⁵ Three alternative definitions of GPS were investigated: (1) eliminating a middle group near the midpoint to maximize high- and low-GPS, with thirty-three (34 percent) low-GPS participants answering no more than three questions correctly and thirty-six (37 percent) high-GPS participants answering six or more GPS questions correctly; (2) splitting so that one-third (thirty-three) fell into a low-GPS group with scores of three or less, and two-thirds (sixty-four participants) were classified as having high-GPS with scores of four or greater; and (3) splitting so that two-thirds (sixty-one participants) were classified as low-GPS with five or fewer questions answered correctly, and thirty-six (one-third), with more than five questions correct, were classified as high-GPS. Results were consistent across this sensitivity analysis.

(3) a short description of each division's target market and strategy, and (4) the scorecards of each division manager, each containing sixteen performance measures and target performance for each manager on each measure. Following the two scorecards, participants were presented with their task: to assume the role of the CFO in evaluating the performance of the two division managers. The two division scorecards were then presented again, this time also giving the managers' actual performance on each item and the percentage difference in performance from target versus actual.

At this point in the case, the first experimental manipulation appeared. One-half of participants were asked to evaluate performance for each performance measure (i.e., 16 performance evaluations for RadWear and 16 for WorkWear), on a scale from zero to one hundred (the same scale used for the overall evaluation), prior to making an overall performance evaluation. The other half of participants was instructed only to indicate an overall performance evaluation (all other information was identical). Participants were given two evaluation scales ranging from zero ("Reassign") to one hundred ("Excellent"), one for each division manager, and asked to indicate their overall performance evaluations.

After making the overall performance evaluations, information was requested from participants regarding their decision process in determining their evaluations. Then a series of seven manipulation check and control questions was presented. Demographic data and work experience information was also requested to gather participants' gender, age, area of emphasis in graduate work, number of years of full-time work experience, discipline area of full-time work experience, frequency making performance evaluations, and the degree of familiarity with the Balanced Scorecard methodology. The experimental case concluded with the nine GPS questions, followed by a self-report of participants' ability to maintain concentration throughout

the case and motivation to correctly complete the case. The task was designed to be administered in a classroom setting during a class meeting and take approximately forty-five minutes to complete.

Control Variables

Participants' attention and motivation were controlled by asking for self-reported ratings of each. Knowledge of and experience using the BSC were controlled by conducting the experiment as part of a classroom case prior to introduction to the BSC topic, and by gathering data from the participants about prior work experience (including the number of years of full-time work experience), and prior experience with performance evaluations and with the BSC.

Dependent Variable

The dependent variable was measured using a 101-point scale from zero to one hundred, with seven descriptive labels ranging from "Unacceptable" to "Excellent." The dependent variable is the absolute value of the difference in the evaluations of the two division managers (RadWear score - WorkWear score). Since the sum of excess performance across all measures, common and unique, is approximately the same for both divisions, a reduction in the common-measures bias and incorporation of more BSC measures, both common and unique, is shown as the difference in divisional evaluations approaches zero.

The observed range of the dependent variable, DIFF, was from zero to twenty, with a mean (standard deviation) of 5.34 (5.59) and a median of 3.5. The first (third) quartile of DIFF was 0.0 (10.0). Statistical results of the tests of the hypotheses are presented in the following section.

IV. RESULTS

The hypotheses were tested with ANOVA to examine the effects of the two factors, BEFORE and GPS, on the absolute value of the difference in the scores assigned to the division managers. The results appear in Table 1. The ANOVA reveals a significant factor ($p < 0.01$) for GPS, but not for the main effect of BEFORE ($p > 0.40$). Regressing BEFORE on DIFF yields a beta weight on the BEFORE variable of -1.66, which is in the predicted direction, since it indicates evaluating the individual measures prior to making an overall evaluation of the managers did tend to decrease the difference in the two scores. (Recall, a greater positive difference indicates using fewer performance measures, as in a common-measures bias.) However, the effect is not significant ($p = 0.15$). Thus, H1 is not supported, and we fail to conclude that requiring an increase of cognitive effort to evaluate performance on unique as well as common measures results in a decrease in reliance on common measures in holistic performance evaluations.

[Insert Table 1 here]

In a regression of GPS on DIFF, the beta weight for GPS is -3.29 ($p < 0.01$), indicating higher problem-solving ability resulted in a smaller difference in the managers' scores. Additionally, GPSSUM and DIFF have a correlation coefficient of -0.30 ($p < 0.01$). These results are robust across the three alternative classifications of high-low GPS. Therefore, the results support the H2, and we conclude evaluators with higher general problem solving ability effectively use more of the unique information contained in Balanced Scorecards.

We controlled for participants' rating of their familiarity with the BSC by including this variable in the ANOVA as a covariate. The added covariate was marginally significant in the ANOVA ($p = 0.08$) and the other results remained unchanged.

Manipulation Checks

Three questions were included to provide information about participants' attention to the task, each measured on a scale from -5.0 ("strongly disagree") to 5.0 ("strongly agree"): "The two divisions, RadWear and WorkWear, were targeting the same markets;" "The two divisions, RadWear and WorkWear, used some different performance measures;" and "It was appropriate for the two divisions, RadWear and WorkWear, to employ some different performance measures." The means of the 97 responses to the three questions were -3.52, 2.61, and 3.69, respectively, which were all significantly different from the scale midpoint of 0.0 ($p < 0.01$). (See Table 2 for descriptive statistics on all variables of interest included in the study.) Eleven participants agreed or were neutral that the two divisions were targeting the same markets; eight disagreed or were neutral regarding whether the divisions employed some different performance measures, and one participant disagreed that the two divisions should use different performance measures. Removing these cases, either separately by statement or combined for all three statements, makes no qualitative changes in the results reported in Table 1.

[Insert Table 2 here]

Supplemental Analysis

We investigated whether the participants who responded with small differences between the managers' evaluations, i.e., those who did not exhibit common-measures bias, varied

systematically from the other participants. First, a new variable, SMALLDIF, was defined to indicate an absolute value of the difference in scores between 0.0 and 5.0. Fifty-eight participants (60%) exhibited small differences defined in this manner.

We cross-tabulated the dummy variable SMALLDIF with BEFORE, expecting more instances of small differences to coincide with completing the separate evaluation task of the individual measures before completing the overall evaluations of the managers. This cross-tabulation appears in Panel A of Table 3. Although the frequency of small differences is highest in the before condition, there is not enough evidence to reject the hypothesis of independence for the two variables ($\chi^2 = 2.35$, $p = 0.13$). Alternative definitions of SMALLDIF (at differences of up to 1.0, 3.0, and 7.0) resulted in the same conclusion.

[Insert Table 3 here]

We next considered the frequency of small differences across GPS, predicting that participants with high GPS would be more likely to respond with a small difference between managers. Cross-tabulating these variables (see Panel B of Table 3) shows the highest number of small differences for those participants with high GPS. A chi-square test for independence gives enough evidence to reject the hypothesis that the variables are independent ($\chi^2 = 9.14$, $p < 0.01$). This result is robust across the three alternative definitions of SMALLDIF.

We then investigated whether participants who responded with small differences significantly varied in their responses to seven manipulation check/follow-up questions compared to the participants with large differences in their evaluations, and discovered significant differences in four questions. First, the level of agreement with the statement, “The

two divisions ... were targeting the same markets” was significantly lower (beta = -0.95, $p = 0.02$) for the participants exhibiting small differences in their evaluations. Second, the level of agreement with the statement, “The two divisions ... used some different performance measures” was significantly higher (beta = 1.22, $p < 0.01$) for these participants. Third, the level of agreement with the statement, “The case was very difficult to do” (measured on a scale from -5.0 to 5.0) was significantly lower (beta = -1.04, $p < 0.01$) for the participants with small differences. Finally, the level of agreement with the statement, “The case was very realistic” (also measured on a scale from -5.0 to 5.0) was higher (beta = 0.60, $p = 0.05$) for the participants with small differences in their evaluations. These results are robust across the alternative definitions of SMALLDIF.

In addition to evaluating the participants with small observed differences in their ratings, we investigated the responses on the seven manipulation check/follow-up questions from the participants with large, positive differences in their ratings, i.e., those who did seem to exhibit a common-measures bias. A new dummy variable, COMBIAS, was defined, with a value of one for signed differences in ratings greater than 5.0 (and zero otherwise). Twenty-nine participants (30%) responded with differences greater than 5.0, ranging from 6.0 to 20.0, with a mean of 11.28. These participants varied significantly ($p < 0.05$) from all other participants on three of the seven follow-up questions. First, the level of agreement with the statement, “The two divisions ... were targeting the same markets” was significantly higher (beta = 1.26, $p < 0.01$) for these participants. Second, the level of agreement with the statement, “It was appropriate for the two divisions ... to employ some different performance measures” was significantly lower (beta = -0.65, $p = 0.02$) for the high-positive-difference evaluators. Finally, the level of agreement with the statement, “The case was very realistic” was significantly lower (beta = -0.68, $p = 0.04$)

for these participants. These results are robust when the definition of COMBIAS is changed to include all positive-difference scores.

We were interested in further examination of this COMBIAS group. For ease of analysis, the 97 participants were categorized into only two groups, one for positive difference scores (“GRP1”) and the other for nonpositive difference scores (“GRP2”). This conveniently divided the sample in half, with 49 participants having difference scores ranging from 1.0 to 20.0, and 48 subjects with difference scores of zero and below. Additionally, this biased against results, since the scores surrounding zero are arguably not very different. Eight participants responded with a difference between managers of only 1.0.

Targeting the Same Markets

Overall, study participants disagreed that the two WCS divisions in the case were targeting the same markets. However, the group of participants who rated RadWear’s performance higher (GRP1) disagreed less strongly than the participants (GRP2) who rated the performance of the divisions the same or scored WorkWear’s performance higher (GRP1 mean was -3.1 versus a mean of -3.9 for GRP2, $p = 0.05$). This difference in means is not due to outliers. Six of the forty-nine responses in GRP1 were zero or higher (with positive responses indicating agreement that the divisions were targeting the same markets) versus only three of the forty-eight responses in GRP2. Overall, a similar proportion of participants in each group disagreed with this statement (86 percent of participants in GRP1 responded with a score below zero, versus 92 percent in GRP2), but the pattern of responses to this statement reveals the GRP2 respondents were more adamant about their disagreement. Sixty-three percent of respondents in GRP2 strongly disagreed (response = -5.0) to this question, versus only 37 percent of

respondents in GRP1. This pattern of responses resulted in the higher mean for the GRP1 participants.

Using the Same Performance Measures

Overall, study participants agreed it was appropriate for the two WCS divisions to use different performance measures. However, the GRP2 participants agreed more strongly with this statement (the mean response was 4.0) than their GRP1 counterparts (mean response of 3.4). As in the previous discussion, the pattern of responses to this statement is different between the two groups. Only one participant responded with a nonpositive agreement score (thus, indicating disagreement), and this participant belonged to GRP1. Additionally, 42 percent of GRP2 strongly agreed with this statement (response = 5.0), versus only 20 percent in GRP1. While the GRP1 participants did tend to agree that the use of different performance measures was appropriate, these participants were not as committed, on average, to their response as those in GRP2.

To provide assurance these results were not due to participants' carelessness (e.g., marking a single response for all questions without attending to the items, etc.), the responses of the participants which were opposite of expectation (those responses which resulted in the differences outlined above) were investigated. Of the forty-nine responses in GRP1, seven cases contained one "opposite" response, versus only three of the forty-eight responses in GRP2 (no case had more than one response contrary to expectation).

The differences in responses are also not due to differences in reported levels of concentration or motivation. GRP1 (GRP2) reported a mean concentration level of 6.0 (5.3), and a mean motivation level of 5.8 (5.6). The groups had similar ranges and dispersion for

concentration and motivation scores. The implications of these statistical results are discussed in the following section.

V. DISCUSSION

The first hypothesis investigated whether requiring BSC users to evaluate performance on each individual BSC measure before making an overall performance judgment would result in greater usage of unique performance measures. The results show this manipulation is not sufficient for inducing decision-makers to incorporate all of the measures. One possibility is that a stronger manipulation is needed to increase the expenditure of cognitive effort. Slovic and MacPhillamy (1974) theorized that unique measures are less utilized because evaluators must (1) evaluate the differences between the measures as well as (2) evaluate the relative performances on the measures. Our manipulation required participants to assess (2), but not (1). Requiring BSC users to evaluate (1) as well as (2) might increase the expenditure of cognitive effort prior to the overall evaluation to a sufficient extent to eliminate or substantially reduce overweighting of common measures. This is a question for future research. Perhaps an effort-related explanation for the common-measures bias is correct, but a stronger manipulation of effort is needed.

However, it is also possible that the manipulation used was sufficient for inducing effort, and the lack of significant hypothesized results indicates that an effort-based explanation for the common-measures bias does not hold. Ideally, including a measure of effort expended would be useful in interpreting the results. We did request self-reported levels of concentration and motivation, but the placement of these items at the end of the experiment, immediately following the GPS measure, renders these responses less useful since it combines concentration and

motivation on both the evaluation task and the GPS items. Comments received from participants both in writing on the instrument and informally at the conclusion of the experiment revealed many participants maintained high levels of concentration and motivation until presented with the GPS items. This decline in motivation and concentration then depressed participants' self-ratings. In future research, the placement of either the GPS items or the concentration and motivation self-reports in the instrument should be changed⁶.

The second hypothesis posited an ability effect on the common-measures bias, predicting decision-makers with higher problem-solving ability to be less susceptible to underweighting unique measures. The absolute value of the difference in scores of the higher ability managers was significantly lower (consistent with expectation) from the scores of the judges with lower problem-solving ability. Robust results were shown for an ability effect, despite the possibility that it may actually be another type of ability, rather than general problem-solving ability, which would be a better fit for this task. Additionally, significant results were obtained despite the fact the GPS measure may have been tainted by participants' fatigue or declines in motivation when presented with the GPS measure at the end of the experimental instrument.

Although it is premature to make exaggerated statements regarding the COMBIAS results in the supplemental analysis, these responses do raise questions about the nature of the common-measures bias. Common-measures bias has been viewed as a judgment bias in which the decision-maker overweighs common measures unintentionally or unknowingly (Slovic and MacPhillamy 1974). However, responses of participants in the current study reveal some may have intentionally relied upon the common measures, believing this to be the rational decision strategy. These participants viewed the divisions as targeting the same markets (and thus, may

⁶Time to complete, in minutes, was collected from some participants. This is problematic as a measure of effort since time spent is a function of both effort and ability. In addition, these data were not available for all subjects.

view the use of common measures as being appropriate), and were less committed to the idea that the divisions should use different performance measures. These beliefs could explain the participants' reliance on the common measures, but importantly, by choice, rather than as a result of a judgment bias. This is a very intriguing possibility which should be explored further in additional research.

VI. CONCLUSIONS

This research finds decision-makers' problem-solving ability directly impacts the amount of BSC information used in evaluating the performance of division managers. Evaluators with higher levels of problem-solving ability incorporate more performance measures, both common and unique, in their evaluations, compared to judges with lower problem-solving ability. An effect was hypothesized relating to decision-makers' effort, expecting that participants expending increased effort would incorporate more performance measures in their evaluations. However, the data did not support this hypothesis.

The results should be interpreted with a reasonable amount of caution, since they may not generalize to other settings. The study utilized a small number of participants, which though sufficient for obtaining statistical significance, limits the generalizability of the results. In addition, graduate business students, lacking in business experience, may not produce results representative of managers making performance evaluation decisions in an organizational context. The artificial case setting, necessary for the experimental design, also limits generalizability. Participants did not have prior knowledge of or experience with the firm in the case or with the divisional managers they evaluated. Having only a few pages of information about an organization does not align with the practical and experiential knowledge a "real-

world” decision-maker would have in this context. This feature of the research is useful for control, but creates a limitation in interpreting the results. It is also artificial that both divisions would outperform their targeted goals on every performance measure. However, this was included to control for perceived differences between positive and negative information, and since this design feature has been utilized in prior studies in this research stream, it was maintained in the current study for ease of comparing results with those obtained in prior work.

Despite the limitations inherent in the research design, the results of the study contribute to our knowledge of common-measures bias in important ways. This study is the first to document the role of problem-solving ability in mitigating the common-measures bias, an important finding since it provides information to organizations using the BSC for performance evaluations about a necessary characteristic of superiors making these decisions. Roberts et al. (2002) showed that MBA students demographically similar to those in the current study may lack the knowledge, ability, or experience to appropriately incorporate the unique measures, whereas Executive MBA participants with an average of ten years of work experience did incorporate the unique measures when simply presented with a reminder that all of the measures were important in the evaluation. This study extends our knowledge by pointing to problem-solving ability, rather than simply work experience, as enabling decision-makers to incorporate the unique measures. The participants in the current study were not provided with the explicit reminder given in Roberts et al. (2002), yet those with high problem-solving ability did include the unique performance measures in their judgments.

This research is the first to directly test the effort hypothesis for the common-measures bias, suggested and assumed since Slovic and MacPhillamy (1974). Surprisingly, however, significant results were not found for the manipulation of effort. It is possible that the

manipulation of increased effort used in this study, i.e., having half the participants consider each BSC item individually prior to making their overall performance evaluations, was not sufficient for inducing increases in effort. A second possibility is that the manipulation did induce increased effort, but separate cognitive processes take place for the two types of judgments (evaluating sixteen performance measures of a single division manager versus making comparative evaluations to assign overall ratings to the two divisional managers), and decision-makers may abstract a criterion other than “include all the BSC performance measures” when making the comparative evaluation. An alternative decision criterion may be fairness, where decision-makers perceive they can make the fairest evaluation by directly comparing managers on their common measures. If this is the case, the effect is not best understood as a common-measures bias, in which a decision-maker is unintentionally drawn to the common-measures because they are more salient, more memorable, or easier to process, but rather may be seen as the judge willfully relying on the common measures to pursue a chosen, rational decision strategy. This is an important distinction to investigate further in future research. It is premature to conclude, on the basis of these results alone, that the effort-based hypothesis is incorrect, but the fact that this research raises the possibility is an interesting and important contribution.

The results also revealed another possible alternative for the common-measures bias in the BSC performance evaluation context. Those participants with large, positive differences in their evaluations (those strongly favoring the division with superior performance on the common measures) were less convinced the divisions were targeting different markets and that they should use different performance measures. This is additional evidence pointing to an intentional focus on the common measures, and not merely a simplifying strategy to reduce cognitive strain, as has been assumed.

Further research is needed to draw conclusions regarding the nature of the effort-based explanation for the common-measures bias to determine whether cognitive strain and an effort-related effect are underlying the bias. More research effort should also be undertaken to better understand those decision-makers who are influenced by the common-measures bias versus those who are not. One piece of that understanding, the problem-solving ability of the judges, was provided by this study. Additionally, the very nature of the common-measures bias is not yet understood. As this effect has implications for contexts beyond performance evaluation and the BSC, it is worthwhile to engage in more research to fully understand the bias, and ultimately, the conditions necessary for mitigating its effects.

REFERENCES

- Ackerman, P. L. 1984. A theoretical and empirical investigation of individual differences in learning: A synthesis of cognitive ability and information processing perspectives. Unpublished doctoral dissertation, University of Illinois, Urbana.
- , 1986. Individual differences in information processing: An investigation of intellectual abilities and task performance during practice. *Intelligence* 10: 101-139.
- , 1987. Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin* 102: 3-27.
- , 1988. Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General* 117: 288-318.
- Banker, R. D., H. Chang, and M. J. Pizzini. 2004. The balanced scorecard: Judgmental effects of performance measures linked to strategy. *The Accounting Review* 79 (1): 1-23.
- Bonner, S. E., and B. Lewis. 1990. Determinants of auditor expertise. *Journal of Accounting Research, Supplement*: 1-20.
- , J. S. Davis, and B. R. Jackson. 1992. Expertise in corporate tax planning: The issue identification stage. *Journal of Accounting Research, Supplement*: 1-28.
- Dilla, W. N. and P. J. Steinbart. 2005. Relative weighting of common and unique balanced scorecard measures by knowledgeable decision makers. *Behavioral Research in Accounting* 17: 43-54.
- Kanfer, R., and P. L. Ackerman. 1989. Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition [Monograph]. *Journal of Applied Psychology* 74: 657-690.
- , -----, T. C. Murtha, B. Dugdale, and L. Nelson. 1994. Goal setting, conditions of practice, and task performance: A resource allocation perspective. *Journal of Applied Psychology* 79 (6): 826-835.
- Kaplan, R., and D. Norton. 1996. *The Balanced Scorecard*. Boston: Harvard Business School Press.
- and -----, 2001. *The Strategy-Focused Organization*. Boston: Harvard Business School Press.
- Kennedy, J. 1993. Debiasing audit judgment with accountability: A framework and experimental results. *Journal of Accounting Research* 31 (Autumn): 231-245.

- , 1995. Debiasing the curse of knowledge in audit judgment. *The Accounting Review* 70 (2): 249-273.
- Kivetz, R., and I. Simonson. 2000. The effects of incomplete information on consumer choice. *Journal of Marketing Research* 37 (November): 427-448.
- Krumwiede, K. R., M. R. Swain, and D. L. Eggett. 2002. The effects of task outcome feedback and broad domain evaluation experience on the use of unique scorecard measures. Working paper, Brigham Young University.
- , -----, and T. V. Eaton. 2004. The effects of strategic linkage and target focus on the use of financial and nonfinancial measures in Balanced Scorecard evaluations. Working paper, Boise State University.
- Libby, T., S. Salterio, and A. Webb. 2004. The balanced scorecard: The effects of assurance and process accountability on managerial judgment. *The Accounting Review* 79 (4): 1075-1094.
- Lipe, M. G., and S. E. Salterio. 2000. The balanced scorecard: Judgmental effects of common and unique performance measures. *The Accounting Review* 75 (3): 283-298.
- Markman, A. B., and D. L. Medin. 1995. Similarity and alignment in choice. *Organizational Behavior and Human Decision Processes* 63 (2): 117-130.
- Norman, D. A., and D. B. Bobrow. 1975. On data-limited and resource-limited processes. *Cognitive Psychology* 7: 44-64.
- Roberts, M. L., T. L. Albright, and A. R. Hibbets. 2002. Utilization of unique measures in the balanced scorecard: Effects of awareness and experience. Working paper, The University of Alabama.
- , -----, and -----, 2004. Debiasing balanced scorecard evaluations. *Behavioral Research in Accounting* 16: 75-88.
- Slovic, P., and D. MacPhillamy. 1974. Dimensional commensurability and cue utilization in comparative judgment. *Organizational Behavior and Human Performance* 11: 172-194.
- Tan, H.-T., and A. Kao. 1999. Accountability effects on auditors' performance: The influence of knowledge, problem-solving ability, and task complexity. *Journal of Accounting Research* 37 (1): 209-223.
- Zhang, S., and A. B. Markman. 1998. Overcoming the early entrant advantage: The role of alignable and nonalignable differences. *Journal of Marketing Research* 35 (November): 413-426.

TABLE 1
INFLUENCE OF EFFORT AND GPS ON OVERALL PERFORMANCE
EVALUATIONS USING THE BSC: RESULTS OF A 2 X 2 ANOVA

<i>Variable</i>	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>
BEFORE	1	19.407	19.407	0.666	0.416
GPS	1	216.865	216.865	7.444	0.008
BEFORE x GPS	1	5.442	5.442	0.187	0.667
Error	93	2709.260	29.132		

BEFORE is a dummy (1/0) variable indicating whether participants explicitly evaluated performance on individual BSC items before (1) or after (0) making overall performance evaluations of the divisional managers.

GPS is the sum of correct answers on the general problem-solving ability measure developed by Bonner and Lewis (1990) and Bonner et al. (1992).

The dependent variable is the absolute value of the difference in participants' evaluations of the two divisional managers (RadWear score – WorkWear score).

TABLE 2
DESCRIPTIVE STATISTICS OF SELECTED VARIABLES

<u>VARIABLE</u>	<u>N</u>	<u>MIN</u>	<u>MAX</u>	<u>MEAN</u>	<u>STD DEV</u>
DIFF	97	0	20	5.39	5.59
GPS	97	1	9	4.58	2.04
AGE	97	21	49	25.61	6.00
“The measures were usefully categorized”	97	-3	5	2.70	1.29
“The two divisions were targeting the same markets”	97	-5	4	-3.52	2.00
“The two divisions used some different performance measures”	97	-5	5	2.61	1.75
“It was appropriate for the two divisions to employ some different performance measures”	97	-3	5	3.69	1.26
“The case was very easy to understand”	97	-3	5	2.87	1.80
“The case was very difficult to do”	97	-5	3	-2.57	1.93
“The case was very realistic”	97	-3	5	2.74	1.48
Familiarity with the BSC	88	-5	5	0.41	2.76
Work Experience (years)	96	0	20	2.79	4.68
Concentration	96	1	10	5.66	2.10
Motivation	96	0	10	5.74	2.05

DIFF is the absolute value of the difference in participants' evaluations of the two divisional managers (RadWear score – WorkWear score).

GPS is the sum of correct answers on the general problem-solving ability measure developed by Bonner and Lewis (1990) and Bonner et al. (1992).

TABLE 3
SUPPLEMENTAL ANALYSIS: COMPARING PARTICIPANTS WITH SMALL
DIFFERENCES IN EVALUATIONS ACROSS EXPERIMENTAL GROUPS

Panel A: Cross-tabulation of observed small differences in divisional performance evaluations and timing of rating individual BSC measures

		After	Before	Total
SMALLDIF	0	23	16	39
	1	25	33	58
Total		48	49	97
		$\chi^2 = 2.35$	($p > 0.10$)	

Panel B: Cross-tabulation of observed small differences in divisional performance evaluations and general problem-solving ability

		Low GPS	High GPS	Total
SMALLDIF	0	27	12	39
	1	22	36	58
Total		49	48	97
		$\chi^2 = 9.14$	($p < 0.01$)	

SMALLDIF is defined as a difference in evaluations (RadWear score – WorkWear score) ranging from -5.0 to 5.0.