MITIGATING COMMON MEASURES BIAS: CAN TRAINING AND ORGANIZATIONAL DESIGN ALLEVIATE MANAGERIAL BIAS?

By

Luke Weiler

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Business Administration - Accounting - Doctor of Philosophy

2020

ABSTRACT

MITIGATING COMMON MEASURES BIAS: CAN TRAINING AND ORGANIZATIONAL DESIGN ALLEVIATE MANAGERIAL BIAS?

By

Luke Weiler

Accounting research has established that when presented with common and unique performance measures about different divisions or managers, decision makers underweight unique information and overweight common information. This "common measures bias" leads to performance strategies that can be inconsistent with the strategy of the firm. I examine how firm strategy (as exhibited through organizational design) influences the common measures bias. While training can influence common measures bias, which *aspects* of training and *how* training influences this bias has not been explored. I experimentally investigate how organizational design and training influence common measures bias. I find that while organizational design appears to have minimal effect on its own, there is an interactive effect of organizational design and training. Additionally, training that emphasizes the inclusion of all metrics increases the weights placed on non-financial metrics. While the training increases the weights on non-financial metrics, participants continue to rely on financial metrics when making their performance evaluation ratings.

ACKNOWLEDGMENTS

I am grateful for the support and guidance of my dissertation committee, Ranjani Krishnan (Chair), Martin Holzhacker, Wayne Nesbitt, and Cheri Speier-Pero. I appreciate the efforts of Jing Kong, Hari Ramasubramanian, Joanna Shaw, and Harlow Lochramirez in running the experiment. I am grateful for the financial support provided by the Broad College of Business. All errors or omissions are my own.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

CEO	CHIEF EXECUTIVE OFFICER
BSC	BALANCED SCORECARD
ANOVA	ANALYSIS OF VARIANCE
GPA	GRADE POINT AVERAGE
SAT	SCHOLASTIC APTITUDE TEST

INTRODUCTION

A survey of the Fortune 1000 companies suggests that approximately 50% use the Balanced Scorecard (Calabro 2001, Crabtree and DeBusk 2008). These companies use the Balanced Scorecard for a variety of purposes (Wiersma 2009). However, while use of the Balanced Scorecard is common, firms continue to rely primarily on financial criteria when determining compensation (Epstein and Roy 2005). An analysis of Fortune's Most Admired Companies list shows that while 100% contained financial performance criteria in Chief Executive Officer (CEO) compensation, only 42% contained customer criteria, 13.5% contained business process criteria, and 10% had employee growth criteria (Epstein and Roy 2005). This primacy of financial information and lack of incorporating non-financial information throughout the business is in stark contrast with the stated goals of the development of the Balanced Scorecard (Kaplan and Norton 1992).

The ability to only incorporate certain information into decisions mirrors the common measures bias found previously. Lipe and Salterio (2000) identified that supervisors focus on common metrics when making performance evaluation decisions and exclude the information value of unique metrics. The psychological mechanism that drives common measures bias is generally understood to be cognitive strain associated with complex decision-making tasks (Payne 1976; Slovic and MacPhillamy 1974; Markman and Medin 1995; Hibbets et al. 2006). Common measure bias leads to performance evaluations that can be inconsistent with the strategy of the company. Tools and strategies that can ease this strain include training (Dilla and Steinbart 2005; Libby et al. 2004), providing strategy information (Banker et al. 2004; Humphrey and Trotman 2011), providing feedback (Krumwiede et al. 2013), and information ordering (Kaplan and Wisner 2009; Lipe and Salterio 2002; Roberts et al. 2004). Frameworks

such as the Balanced Scorecard (Kaplan and Norton 1992) can also assist managers to focus on strategy and identify the drivers of future performance.

This paper examines whether organizational design and training can mitigate the common measures bias. Cognitive biases, such as the common measures bias, typically occur when decision making takes place as an intuitive process as opposed to an analytical process (Croskerry et al. 2013). Decision making processes rely more on intuitive processes and the heuristics inherent in those processes when decisions are complex (Croskerry et al. 2013; Frederiks et al. 2015; Schwenk 1988). While heuristics provide shortcuts to ease cognitive strain, they also introduce the possibility for suboptimal decisions through the introduction of biases (Mosier et al. 1998). Decision aids can provide cues and prompts that can ease processing by pointing to specific information and strategies to use in decision making (Croskerry et al. 2013; Mosier et al. 1998; Frederiks et al. 2015).

While prior research has investigated how providing divisional strategy information influences performance evaluations (Banker et al. 2004; Humphrey and Trotman 2011), strategy can also influence performance evaluations through organizational design. Corporate strategies determine the extent of interdependence in organizational design (Gupta and Govindarajan 1986; Pitts 1977; Chenhall and Morris 1986). Firms can either be designed to emphasize competition or cooperation between divisions (Hill et al. 1992; Gupta and Govindarajan 1986). While cooperative firms attempt to gain benefits of economies of scope through their synergistic strategy, competitive firms attempt to gain benefits through efficiencies of internal governance mechanisms. The difference in level of interdependence further determines the choice of performance evaluation metrics (Gupta and Govindarajan 1986; Chenhall and Morris 1986). However, the relationship between organizational design and performance metrics is much more

complex. While organizational design influences the choice and importance of specific performance metrics, it is unclear whether supervisors understand and use this information in their performance evaluation decisions. I test whether organizational design, as articulated through information about competitive or cooperate design, mitigates the common measures bias.

While Dilla and Steinbart (2005) found that Balanced Scorecard (BSC) training can mitigate some of the common measures bias, their research does not test which aspects of training mitigate the common measures bias. The BSC philosophy emphasizes that non-financial metrics have informational value about performance (Kaplan and Norton 1992). As such, using both financial and non-financial metrics as well as both common and unique metrics in performance evaluation decisions would theoretically provide the best performance ratings. BSC training typically emphasizes the importance of relying on all performance metrics when making decisions. I test whether a three minute BSC training session changes the self-reported weights placed on metrics. Additionally, I test whether this short training mitigates the common measures bias.

I test my predictions through a 2 x 2 x 2 experiment. The first factor is whether common or unique metrics have higher above-target performance. This factor allows me to directly compare my results to those of Lipe and Salterio (2000). The second factor is whether the firm's organizational design is competitive or cooperative. This factor allows me to interpret whether participants understand that firm strategy leads to differential predictions on how metrics should be used in performance evaluations. The third factor is whether BSC training is present or absent. Specifically, I test how training that emphasizes the importance of incorporating all metrics into performance evaluations changes the weights placed on metrics and whether these

weight changes lead to changes in the performance ratings. I find that participants under both competitive and cooperative organizational designs exhibit common measures bias. Further, I find that training does appear to minimally mitigate the common measures bias for participants in the competitive organizational design setting. However, Analysis of Variance (ANOVA) for all participants does not show training as a significant determinant of changes in performance ratings. Additional analysis shows that participants change the importance they place on performance metrics after training. Specifically, the self-reported weights placed on non-financial metrics increase. However, even though the participants report that they increased reliance on non-financial metrics when making ratings decisions, the only significant determinant of ratings changes was the weight placed on financial metrics. So while basic training that emphasizes the importance of all metrics changes the weights placed on metrics, participants relied on only the financial metrics when making their decisions.

A second experiment was conducted to further investigate the impact increased levels of training (as defined in Bloom's Taxonomy) may have on mitigating the common measures bias. The second experiment was designed where the training consisted of more experiential training with a hands-on exercise. This simulation is linked to the knowledge, comprehension, application, and analysis levels of Bloom's Taxonomy while the training in the first experiment only addressed the knowledge level (Miller et al. 2014). Results show that increased levels of training continue to impact the weights placed on non-financial metrics. However, participants continued to rely more on financial measures when making their performance evaluation decisions.

My study helps contribute to our understanding of the common measures bias in a number of ways. First, the type and duration of training appears to matter when attempting to

mitigate cognitive biases. While short training may appear to influence how decisions are made, these changes may only be cosmetic and not actually change the underlying decision process. Experiential simulation-based training may lead to greater internalization by participants and integration of the training into their decisions. Further investigation of different aspects of training and the duration of training could yield further insights into their effects on mitigating cognitive biases. Second, this paper addresses the relative use of financial and non-financial metrics in performance evaluations. While participants recognized and stated that non-financial metrics were important to their decisions, the reliance on financial metrics for performance evaluations appears to be difficult to overcome. It may be that these financial metrics provide a familiar beacon to focus on when placed in complex decision tasks. Thus, the overweighting of financial measures and the underweighting of unique measures appears to be stubborn biases that will require more extensive training tools to mitigate.

RESEARCH SETTING AND HYPOTHESIS DEVELOPMENT

Common Measures Bias

Slovic and MacPhillamy (1974) suggest that people use common and unique information differently when making decisions. They asked participants to estimate a student's Grade Point Average (GPA) from scores on common dimensions (English skills) and unique dimensions (Quantitative aptitude for student A and Need to Achieve Success for student B). They found that participants weighted the common information more heavily. Feedback, monetary incentives, and whether the information for student A and B were given together or separately did not change the decisions by participants.

Lipe and Salterio (2000) use the logic of Slovic and MacPhillamy (1974) within the BSC setting and find participants differentially use common and unique performance metrics. Participants focus their attention on the common metrics, even when the unique metrics provide more information about performance. This phenomenon is called common measures bias, where supervisors who have multiple subordinates fixate on performance metrics that are common between subordinates to the exclusion of unique performance metrics that are tailored specifically to the strategy of the specific divisions (Lipe and Salterio 2000; Slovic and MacPhillamy 1974).

The psychological mechanism that drives common measures bias is generally understood to be cognitive strain associated with complex decision-making tasks (Payne 1976; Slovic and MacPhillamy 1974; Markman and Medin 1995; Hibbets et al. 2006). Payne (1976) argues that people employ different information processing strategies based on task complexity. For simpler decisions, participant employed all relevant information in their decision process. However, for more complex tasks, participants first employed strategies to eliminate alternatives as quickly as

possible on the basis of only a limited amount of information search and evaluation. Markman and Medin (1995) investigate the justifications for decisions between alternative characteristics of products. They find that justifications systematically favored comparable over noncomparable properties. Their findings support the premise that individuals focus on those characteristics that are comparable when asked to make higher level cognitive processing decisions.

Subsequent work to Lipe and Salterio (2000) has found certain conditions under which common measures bias is diminished or mitigated. First, the common measures bias can be partially mitigated when strategy information is provided and when participants are made aware that performance metrics are linked to the strategy (Banker et al. 2004; Humphrey and Trotman 2011). Banker et al. (2004) show that strategically linked metrics influence ratings more than non-strategically linked metrics, but only when participants are provided detailed information about divisional strategies. However, Banker et al. (2004) also find evidence that supports the common measures bias. While participants rely more on strategically linked metrics, they continue to rely more on common metrics than unique metrics. Humphrey and Trotman (2011) expand Banker et al. (2004) to show that divisions with fully linked metrics and detailed strategy information can combine to mitigate the common measures bias. This mitigation only occurs when both fully linked metrics and detailed divisional strategy information is provided. If only one is provided, the common measures bias remains. Information on both strategically linked metrics and divisional strategies could decrease cognitive strain by providing information on the reliability and relevance of the performance metrics. In this way, participants don't have to identify the importance and can simply interpret the information value of each metric.

Second, feedback and experience conducting performance evaluations can partially mitigate the common measures bias. While Slovic and MacPhillamy (1974) identify that

notifying participants of "correct answer" feedback did not eliminate the common measure bias, Krumwiede et al. (2013) identify that providing round-by-round feedback on which metrics are leading indicators of future performance does reduce the common measures bias. Additionally, they find that participants with significant experience performing supervisory evaluations are less likely to exhibit common measures bias. Taken together, the findings of Krumwiede et al. (2013) suggest that learning effects can mitigate the cognitive strain associated with common measures bias.

Third, the presentation structure of metrics can affect the common measures bias. Lipe and Salterio (2002) find that whether metrics are categorized into BSC categories or left "freeform" impacts evaluations. Similarly, whether above-target metrics are concentrated within one BSC category or are scattered among the categories can impact evaluations (Lipe and Salterio 2002). Cognitive search heuristics, specifically the "divide and conquer" strategy leads to these differences in evaluations. The divide and conquer strategy identifies that individuals choose to process packets of information when given large, complex tasks. Structures that allow easy processing of the information will impact the overall evaluations by focusing the heuristics within those structures. For example, providing above-target performance metrics all within one BSC category makes it easier to identify differences between divisions. Spreading those above-target metrics across all four BSC categories makes it more difficult for participants to identify the patterns within the metrics. Roberts et al. (2004) employ an alternative structural device whereby participants evaluated each of the 16 metrics in a disaggregated manner and then mechanically aggregated the overall results. Participants were allowed to subjectively alter the mechanical aggregated results to arrive at a final evaluation. They find that providing a cognitive tool in the mechanical aggregation can mitigate the common measures bias.

Performance Measurement and the Balanced Scorecard

The Balanced Scorecard is a performance measurement system under which the common measures bias can be investigated. While the common measures bias specifically describes the use of metrics based on their characteristics, we know in accounting that performance metrics are part of a larger performance measurement system in place at firms. The performance measurement system typically used in the accounting literature investigating the common measures bias is the BSC.

Kaplan and Norton (1992) describe the concepts of the BSC as a set of metrics that give managers a comprehensive view of the business. The BSC was a tool that was developed in response to the over-reliance on financial metrics that the authors saw in their interactions with businesses. The BSC includes both financial measures that provide the results of actions already taken as well as non-financial measures such as customer satisfaction, internal processes, and organizational innovation and growth that are the drivers of future financial performance (Ittner and Larcker 1998).

Key to the design of the BSC is the link to strategy. Kaplan and Norton (1992) argue that traditional performance measurement systems were designed primarily with an eye to control. These systems specify particular actions desired from employees and then measure whether employees acted as specified. BSC, on the other hand, sets goals linked to strategy and allows whatever actions are necessary to achieve those goals. Kaplan and Norton (1996) argue that a properly constructed BSC should tell the story of the business unit's strategy. Additionally, BSC are typically designed from the top down. This allows cascading goals to be tailored to each subsequent subordinate level's particular circumstances. While the firm may have a goal of growth, two subordinate divisions may take different steps to achieve this goal of growth given

their operational differences. Supervisors rely more on strategically linked performance metrics in making decisions when they know those measures are linked to the strategy of the business unit (Banker et al. 2004). Humphreys and Trotman (2011) found that both strategically linked metrics and knowledge of the unit's strategy are needed to change the decision making. Having strategically linked metrics may also lead to the prevalence of unique metrics within divisions as divisions can have unique strategies to achieve their goals.

One key assumption of the BSC is that nonfinancial performance metrics provide value beyond that of the financial metrics (Kaplan and Norton 1992). Extant research has found that nonfinancial performance metrics are leading indicators of future financial performance (Ittner and Larcker 1998; Dikolli and Sedatole 2007). The relationships between current nonfinancial metrics and future financial outcomes are complex, with various characteristics of information content affecting the relationship (Dikolli and Sedatole 2007).

The difference between common and unique metrics that is the underlying structure of the common measures bias is related, but distinct, from the differences between the financial and nonfinancial metric structure of the BSC. Common metrics are measured and designed through shared techniques. Unique metrics are tailored to the needs of specific divisions (Hoque 2004). While financial and nonfinancial metrics can be both common and unique, financial metrics are more likely to be common metrics in firms for two reasons. First, financial metrics are typically numerical in nature. This numerical nature allows these metrics to be more easily comparable and more easily designed from shared techniques. Nonfinancial metrics can be qualitative in nature. This qualitative nature is more difficult to develop from shared techniques and definitions. Second, financial metrics are historically aggregated into the financial statements while nonfinancial metrics are operational and reserved at the local level. While there have been

movements to incorporate nonfinancial metrics into traditional financial statements and other public disclosures, most public disclosures remain financial in form. This aggregation not only allows for the dissemination of the shared techniques of measurement, but also reinforces the familiarity of those financial metrics.

Strategy, Organizational Design, and Performance

Zimmerman (2011) defines firm architecture as a "three-legged stool" of performance measurement, performance reward and punishment, and decision rights allocation. These three activities are the administrative devices within the firm. For this paper, I focus on the two "legs" of performance measurement and decision rights allocation. Performance measurement was discussed in the previous sections on the common measures bias and the use of the BSC performance measurement system. I manipulate decision rights allocation using the organizational design of the firm. Organizational design includes the traditional centralization/decentralization decision discussed by Jensen and Meckling (1992) but can also include other decisions by management that impacts the relational structures within the firm and the extent of decision-making autonomy that resides at the local level.

Hoque (2004) investigates the relationship between management strategic choice and performance evaluation using survey data. He finds that strategy impacts the organization's performance indirectly through choice of non-financial performance metrics. In other words, as strategies are implemented, firms are more likely to choose non-financial performance metrics to track adherence to goals. In turn, the use of non-financial performance metrics is linked to organizational performance. This supports the notion that strategic initiatives gain more explanatory power through non-financial metrics.

Govindarajan and Gupta (1985) investigate the linkages between firm strategy and control system design. They identify the most critical strategic issues for divisions within firms as being the divisional mission, competitive posture, and the extent and nature of linkages with other divisions within the firm; however, their study only focuses on the strategic mission of the divisions. Hill et al. (1992) argue that diversification strategies (i.e. the extent and nature of linkages with other divisions within the firm) are associated with different sets of economic benefits. They find that appropriate fit between strategy, structure, and control system is associated with superior performance.

> *"Firms attempting to realize economies of scope perform better if their organizational arrangements stress cooperation between business units, while firms attempting to realize economic benefits from efficient internal governance perform better if their organizational arrangements stress competition between business units. (Hill et al. 1992)"*

Because firms design their performance evaluation system to meet their particular strategy (Hoque 2004), there is a high likelihood of using unique metrics for divisions that have different strategies. This use of unique metrics aligns with the concept of internal governance where the unique metrics provide more information about the internal control environment within the division than would common metrics. Alternatively, Hill et al. (1992) argue that in order for firms to realize the benefits of efficient internal governance, they must stress competition. Since common metrics are much easier for supervisors to compare across divisions, focusing on common metrics would achieve this task better than focusing on unique metrics. Thus both common and unique metrics provide benefits for divisions operating as competitors and would both be used in an ideal situation.

Alternatively, when firms are attempting to gain the benefits from economies of scope, Hill et al. (1992) posit that cooperation is preferable. An emphasis on cooperation in employees' performance evaluations would mean avoiding a focus on common metrics because common metrics are directly comparable between divisions and would instill a competitive culture. Thus, while firms may still gain benefits from unique metrics that provide information on specific strategies and initiatives, they would ideally avoid common measures under cooperative designs.

While ideal situations for both competitive and cooperative structures provide different assumptions on the use of performance metrics, the common measures bias is difficult to overcome. Slovic and MacPhillamy (1974) found that the common measures bias persists despite feedback and monetary incentives. The accounting literature on common measures bias has found that only specific pieces and combinations of information mitigate a portion of the common measures bias. Even with the manipulations within these studies, participants continue to rely on common metrics when unique metrics provide more information. Thus, the common measures bias appears to be persistent. Therefore, I predict:

H1: Participants presented with competitive organizational designs will continue to exhibit the common measures bias.

H2: Participants presented with cooperative organizational designs will continue to exhibit the common measures bias.

Training's Impact on Common Measures Bias

Dilla and Steinbart (2005) investigate the affect BSC training can have on the common measures bias. They find that participants knowledgeable of the concepts and underlying rationale of the BSC incorporated both common and unique metrics into their performance evaluations. However, their training was given in a classroom setting and there was no control

group in their study, so questions remain as to whether training in a controlled experimental design can mitigate the common measures bias.

What type of training would be best to mitigate the common measures bias remains a question for inquiry. On one hand, it would appear that direct training on the common measures bias would be best to mitigate the bias. However, Slovic and MacPhillamy (1974) found that when participants were directly told to not weight common measures higher than unique measures, the common measures bias remained. So it appears that a direct approach to training may not be most effective. An alternative training could focus on the system by which performance metrics are presented, namely the BSC. If participants fully understood and incorporate the BSC concept of using all metrics for decision making, the common measures bias would be mitigated.

Literature has not investigated the interaction between training and organizational design. The basic concept of BSC training is to incorporate all metrics into decision making. This concept aligns with the ideal use of both common and unique metrics from competitive organizational designs. Thus the BSC training helps reinforce this ideal use of metrics provided by the competitive prompt and should move participants away from the common measures bias. In cooperative organizational designs, the ideal use of metrics focuses on unique metrics and avoids common metrics. This goes against the concept of the BSC training to incorporate all metrics and may provide competing prompts to participants. These competing prompts may prevent participants from moving towards the ideal use of metrics under the cooperative organizational design. Therefore, I predict that:

H3: BSC training reduces the common measures bias in competitive, but not cooperative, organizational design settings.

While Dilla and Steinbart (2005) conclude that training mitigates the common measures bias, their design doesn't distinguish how this is accomplished. On one hand, it may be that BSC training typically emphasizes the importance of incorporating all metrics into decision making as all metrics have incremental information about performance. This is especially true when metrics are strategically linked (Banker et al. 2004; Humphrey and Trotman 2011). BSC training could influence performance evaluations through a mediation effect wherein such training changes the importance that evaluators place to individual metrics in decision making. Alternatively, the BSC is a complex report that includes both differences in types of information (common/unique) and organization of the information (financial and non-financial attributes). Managers with little experience in the BSC may initially resort to simplifying strategies when using the BSC (Dilla and Steinbart 2005). These simplifying processes can lead to "severe and systematic errors" (Tversky and Kahneman 1974) such as the common measures bias (Lipe and Salterio 2000). BSC training may simply increase the familiarity of participants with the structure and presentation of information. Thus, increased exposure to the BSC will have a direct effect on performance evaluation and the mitigation of common measures bias without impacting the weights placed on specific metrics. Therefore, I predict that:

H4: Training has a direct effect on performance evaluations.

H5: The effect of Training on performance evaluations is mediated through the importance (i.e. weights) placed on performance metrics.

METHOD

Overview of Experiment

The experimental design builds on Lipe and Salterio (2000). Participants in the experiment were given a role-playing scenario in which they took the role of a senior executive at WCS Incorporated, a fictional firm specializing in apparel. The general scenario introduces participants to the overall firm and its two divisions – WorkWear, which specializes in business apparel for young professional, and RadWear, which specializes in sports apparel for young adults. The design of this scenario, as outlined in Appendix A, aligns with two strategies companies employ to increase sales. WorkWear utilizes a customer shopping experience strategy similar to Stitch Fix. Personal shopping services are increasing in volume with Amazon[1] and Walmart[2] recently joining the ranks of boutique retailers and high fashion designers. RadWear utilizes a cost-cutting strategy by introducing online customization and direct online sales to customers.

Participants were asked to perform tasks in multiple rounds. In the first round, participants were asked to rate each of the two managers based on provided performance metrics for each manager, allocate capital investment funds between the two divisions, and determine which manager performed best. After making their performance evaluation decisions, participants were asked to rate how important each metric for both divisions were in making their decisions on a 1-7 Likert scale.

Participants were then given a video prompt. Half the participants were provided training on the Balanced Scorecard. The remaining participants were given training on how to use the semicolon. Participants were then given a chance to perform a second round of evaluations

[1] Amazon announced their Personal Shopper Program on July 30, 2019 (Bobb 2019)
[2] Walmart began their Personal Shopper Program in 2018 (Jones 2018)

using the same information set as the first round. The directions for this round specifically made it known that the information was identical to the first round. After making their decisions, participants were again asked to rate the importance of each metric for both divisions on 1-7 Likert scales. Finally, participants completed a post-experiment questionnaire that provided manipulation checks, asked for demographic information, and gathered impressions on task realism, fondness, and understanding.

Subjects

One hundred eleven undergraduate students at a Midwestern United States university served as experimental participants. Three participants were eliminated from the sample for not fully completing the rating task. The students had, on average, 1.25 years of work experience, 92.3% were between 21 and 23 years old, and 53.8% percent were male.

Design and Procedure

The experiment employs a 2 x 2 between subjects design with an additional 2-level within subjects factor (i.e., the complete design is 2 x 2 x 2) as shown in Figure 1. The first factor is the metrics of performance that were better for the WorkWear division. Thus, WorkWear could perform better than RadWear on common metrics or WorkWear could perform better than RadWear on unique metrics. Because WorkWear is always performing better than RadWear, this design may be impacted by how participants view the strategy and metrics of the two divisions. However, analysis of the post-experimental questionnaire reveals no significant differences from participants' ratings of whether divisional strategies were appropriate, whether divisional metrics were aligned with the divisional strategies, or whether divisional metrics were adequate indications of divisional performance. Thus, there is no evidence of confounding results due to preferences in divisions.

The second factor is the overall firm strategy as depicted in the organization's design. Participants were given either the competitive or cooperative prompt in their scenario description (See Appendices B and C). This prompt was meant to highlight whether the two divisions work together to achieve the firm's overall goals or whether they work separately to achieve the firm's goals. The experimental scenario was adjusted from the Lipe and Salterio (2000) scenario to include divisions that would realistically be able to either compete or cooperate. Lipe and Salterio's (2000) two divisions focused on urban teenagers and business uniforms. The adjustment in this experiment was to change the focus of the two divisions to young professional apparel and young adult sportswear. This change allows me to focus on cooperation because the focal customer is in the same market or to focus on competition because the two divisions sell apparel for different activities.

The third factor is whether BSC training was received. Half of the participants for each session were assigned to a room that showed a video on how to use the semi-colon from Khan Academy (https://www.youtube.com/watch?v=41XNKfR56OY) and the other half of the participants were assigned to a room that showed a video on the BSC from Harvard University (https://www.youtube.com/watch?v=biyGxEix5Zs&t=34s). This training prompt is meant to highlight the underlying conceptual framework of the BSC to look beyond just the financial data when making decisions. Thus, the BSC training highlights the importance of looking at a larger portion of the performance metrics. The Khan Academy training was chosen to provide a similar level of information to the participants while providing irrelevant training for the current task.

I chose performance data such that unique and common metrics for WorkWear had the same degree of superiority in performance. Thus, when WorkWear outperformed RadWear on unique metrics, WorkWear unique metrics were 8% above target while RadWear unique metrics

were 2% above target and all common metrics for both divisions were 2% above target. Similarly, when WorkWear outperformed RadWear on common metrics, WorkWear common metrics were 8% above target while RadWear common metrics were 2% above target and all unique metrics were 2% above target. This design accomplishes two things. First, it standardizes the level above target between the common and unique metrics manipulation categories. This allows me to directly compare the division manager ratings without adjustments based on the magnitude of performance above target. Second, because each Balanced Scorecard category has two unique and two common metrics, this standardization of performance above target controls for the total above target performance within each BSC category. This design addresses one concern raised by Cardinaels and van Veen-Dirks (2010) that placement of the above target performance impacts the ratings. By placing consistent above target performance evenly throughout the four Balanced Scorecard categories, participants are given information on the performance differences between the two divisions regardless of which BSC category they focus on for their decisions.

RESULTS

H1 predicts that, while the ideal situation for competitive organizational design would incorporate both common and unique metrics, the common measures bias overrides this ideal situation. Results are shown in Table 6. Panel A reports the ANOVA analysis and Panel B reports the comparison of mean ratings between the conditions for the first round. Panel A shows that type of metric (common/unique) influences the performance evaluation ratings; however organizational design (represented by the Compete variable) is insignificant. Panel B shows that supervisors under competitive structures exhibit common measures bias and only focus on the common metrics when making performance evaluation decisions. The results support the conclusion that participants under the competitive organizational design exhibit common measures bias. This result supports my hypothesis.

H2 predicted that, while cooperative firms ideally focus only on unique metrics, they will exhibit the common measures bias. Results in Table 6, Panel B for the first round show similar results to the competitive firms in that cooperative firms also exhibit common measures bias. This result supports my hypothesis.

Taken together, the results of H1 and H2 indicate participants do not fully incorporate the organizational design prompt into their decision making. More exploration of this result is warranted to explore whether the unintended costs associated with misalignment of strategy and use of performance metrics becomes realized under this scenario. The literature on alignment between strategy and performance evaluation system argues that conducting evaluations on metrics misaligned with strategy may lead to distortion behavior by managers in future periods (Baker 2002). An experiment investigating the effect common measures bias has on managerial effort and decisions in subsequent periods would provide insight into this area.

H3 predicted that BSC training reduces the effects of common measures bias for competitive, but not cooperative, organizational designs. This full model is shown in Figure 1. Table 7, Panel A provides the results of this analysis. As shown, none of the variable of interest in the ANOVA analysis are significant. However, the means analysis provided in Table 7, Panel B indicates the combination of training and competitive organizational design partially mitigates the common measures bias as shown by the significant differences in ratings both when common and unique metrics are favored. Cooperative organizational structures continue to exhibit the common measures bias. Thus, results support my hypothesis for an interactive effect of organizational design and training.

Figure 2 provides the model incorporated into H4 and H5. Table 8 provides a comparison of means between common and unique metrics within the four BSC categories. Results show a couple of interesting patterns. First, as shown in columns 3 and 6 of Table 8, participants rate common metrics higher than unique metrics for financial and customer metrics but unique metrics are rated higher than common metrics for internal business and growth metrics. This may indicate that the common measures bias is located primarily within the financial and customer categories. Since these categories contain the major metrics that participants are likely to encounter from accounting courses, this may indicate that familiarity of participants with certain common metrics may be driving the common measures bias and the persistence of the bias. Second, the mean ratings for all metrics in the financial and customer categories are significantly higher than those in the internal business and growth categories (significance not shown in table). Third, the significant changes in ratings between the first and second rounds are situated primarily in the internal business and growth categories (as shown in the last two columns). This may indicate that a "catching up" aspect is in effect for the metrics

in those categories to bring them more in line with the importance ratings of the metrics in the financial and customer categories. Fourth, the only significant difference between divisions appears for unique metrics in the financial and internal business categories. While this indicates a significant differential in importance of those ratings, participants indicated no significant difference in the ratings between the two divisions in the post-experimental questionnaire.

Hypothesis 4 predicts a direct effect from training to performance evaluations. Hypothesis 5 predicts an indirect effect of training through changes in the importance of metrics in decision making for the performance evaluations. Tables 9 and 10 provide the analysis for both of these hypotheses. Table 9 shows the Structural Equation Modeling mediated model of training on the change in performance ratings differences through changes in the average importance of BSC category metrics for both divisions. As shown, training affects the changes in average ratings for WorkWear's Customer metrics, WorkWear's Internal Business metrics, WorkWear's Innovation metrics, and RadWear's Innovation metrics. However, the only significant mediation variable that affects performance evaluations is WorkWear's Financial metrics. This means that while training causes participants to revise upward their average weights on WorkWear's non-financial metrics and RadWear's Innovation metrics, participants focus solely on WorkWear's financial metrics when making their performance evaluation decision. The total indirect effect for any given mediator is insignificant. This provides partial support for H5. The Total Effects portion of the model shows no direct effect of Training on performance evaluations. Thus, no support is shown for H4.

Table 10 provides a slightly different analysis on the mediated model of training on performance evaluations through the importance ratings of groups of metrics. In Table 10, I categorize the importance ratings into the common/unique metrics for each division. As shown,

the effect of training on the mediators indicates that training increases the average rating for WorkWear's common and unique metrics and RadWear's common metrics. However, none of these mediation variables has an effect on the change in performance ratings difference. Similar to Table 9, none of the total indirect pathways through the mediators are significant. This again provides some support for H5. The Total Effects portion of the model shows no direct effect of Training on performance evaluations. Thus, again there is no support for H4.

The differences between my findings and those of Dilla and Steinbart (2005) that training eliminates the common measures bias could be due to the type of training being given. My training was a three minute video on the BSC philosophy that all metrics provide information valuable for decision making. It seems logical that such a prompt would lead to increase importance ratings on non-financial metrics. However, more robust training, different types of training tools, or more experience with BSC may be needed to incorporate knowledge that the non-financial metrics have informational value into action in performance evaluations.

EXPERIMENT 2

Experiment 2 Design

In order to further investigate the findings surrounding the training manipulation, I designed and ran a second experiment. This experiment has two main deviations from the first experiment. First, I remove the organizational design manipulation. This allows me to focus entirely on the influence of training on the common measures bias.

Second, I change the design of the training manipulation to align with a higher level on Bloom's Taxonomy of Educational Objectives. The second experiment uses experiential learning to move to this higher level of Bloom's Taxonomy. Providing a three minute video, as I did in the first experiment, is associated with knowledge which is the lowest level of Bloom's Taxonomy. In this stage, recall of material presented is the only objective (Stearns and Crespy 1995; Tennyson and Merrill 1971). Providing a three minute video whose primary message is to use all the available metrics in decision making may change how participants weight the metrics, but a deeper understanding of the relationships within the metrics is needed to apply the learning material to new situations.

The second experiment changes the training from a knowledge basis to a higher level - namely comprehension, application, and analysis. The new training consists of printed powerpoint slides discussing the objectives and design of a Balanced Scorecard, a hands-on exercise building a strategy map from a word-bank of strategic objectives and potential metrics, and a five multiple choice question quiz to test understanding and application of the material. Participants in the control group were given a similar structure of educational based powerpoint slides, an exercise, and application questions from a practice Scholastic Aptitude Test (SAT) reading comprehension section. This exercise is designed as a branching decision simulation

with non-dynamic reactions and includes Bloom's Taxonomy levels of knowledge, comprehension, application, and analysis (Miller et al. 2014). Higher levels of understanding are theorized to result from even more intricate and dynamic simulations such as a Monte Carlo simulation.

While Bloom's Taxonomy theorizes that higher levels of education, as gained through simulations, can help integrate information into future decisions, it is unclear whether simulation training can overcome cognitive biases and the common measures bias specifically. Dilla and Steinbart (2005) conclude that training can mitigate some of the common measures bias; however their experimental design doesn't allow for identifying which aspects of training mitigate the bias. Dilla and Steinbart (2005) conducted their training over two class sessions and testing on the topic. This training included developing two BSCs for different businesses as well as readings on the BSC. There are, however, a couple potential issues with designing the training as an in-class exercise and having it over a longer period of time. First, there is no included control group as all participants were from the designated class and all students in the class had the training. Without a control group, the design of the manipulation is much less controlled and it is difficult to rule out alternative explanations.

Second, by teaching BSC over a number of class periods, it is unclear whether the results from Dilla and Steinbart (2005) are due to BSC training alone or a combination of the training and a reflection period. Experiential learning is typically conceived as the combination of "learning by doing" and a reflection process (Frontczak 1998). In fact, Dewey (1938) theorized that primary experience is a reaction to sensory stimulation while secondary experience, or the reflective experience, is what actually constitutes understanding (Hunt 1995; Frontczak 1998). By confounding the training and the reflective process together, we don't actually know which is

driving the results of Dilla and Steinbart (2005). My design tests for the immediate reaction to the stimulus. The timeframes of the experiment don't allow for any practical reflective process on the material presented. So while I test for a reaction to the simulation training, participants may not have completely processed the training and been able to apply it with a level of understanding. I retest H4 and H5 using the pilot of the second experiment.

Experiment 2 Results

Eighty participants were recruited for the second experiment from Mechanical Turk. Results are in Tables 11 – 15. Tables 11 and 12 show the first and second round results, respectively. These initial results do not show significant common measures bias; in fact they show that participants were able to integrate the performance information into their evaluation decisions whether it was located in common or unique metrics. Table 11 Panel B shows that WorkWear division is significantly rated higher when above target performance favors common metrics and when above target performance favors unique metrics. The difference in ratings between the two conditions (favor common vs. favor unique) is not significant. Additionally, Table 11 Panel A does not show *Metrics* as a significant driver of difference in ratings between the two divisions.

Table 12 shows similar results to Table 11 for the second ratings period. As shown in Panel B, participants correctly incorporated information into the performance ratings decisions in each condition. Ratings significantly favored WorkWear under both the scenario where above target performance metrics were imbedded in common metrics and when they were imbedded in unique metrics. Panel A shows neither *Metrics*, *Training*, nor the combination of *Metrics and Training* were significant drivers of the difference in ratings between divisions. The training manipulation which was designed to alleviate common measures bias would be rendered

ineffective when the bias wasn't present to begin with. Untabulated results show participant ratings of the strategy, metrics, and whether metrics were aligned with divisional strategy as not significantly different.

Further investigation of this surprising result is warranted. Common measures bias has previously been found to be nearly universal and resilient. One possible explanation is that the participant population from Mechanical Turk was significantly different than those studied before. Most studies of the common measures bias in the accounting setting use business students. The Mechanical Turk population had 73 out of 80 participants with majors outside of business topics. It is possible that the bias in the accounting setting is due not to an internal mechanism that focuses on all common metrics but only certain well known common metrics. Results in Tables 8 and 15 provide some initial support that this may be the case. These tables show that common metrics are higher rated than unique metrics in financial and customer categories, but not in internal business and growth categories. The financial and customer categories of metrics may be more familiar to business students and what we perceive as the common measures bias is in fact a common financial and common customer measures bias.

Tables 13 and 14 investigate the mediating effect of importance placed on the different metrics. Similar to the results from Table 9, Table 13 finds that experiential training changes the weights placed on metrics, specifically both division's financial and innovation and growth metrics. Interestingly, training appears to shift importance from the financial metrics and towards the non-financial, innovation and growth metrics for each division. However, the only two fully mediated pathways are through the WorkWear division. While participants increased their importance rating on innovation and growth metrics, they relied less on those metrics when making ratings decisions. Alternatively, while participants decreased their importance ratings on

WorkWear's financial metrics, they relied more on those financial metrics when making ratings decisions. This supports findings from experiment one in which participants shifted importance away from financial metrics, but still relied on those financial metrics when making decisions. Table 14 shows no significant mediation results when looking at unique and common metrics in the two divisions.

These results support H5 that training is mitigated through the importance placed on different divisional metrics. However, there remains no support for H4 of a direct effect of training on changes in the ratings differences. While the results appear to provide some evidence in support of H5, further investigations may be necessary. The higher level of training provided only appears to superficially change the importance ratings without changing the action of the participants when making decisions. However, I am unable to conclude whether training was effective at mitigating the common measures bias in the second experiment as participants did not exhibit this bias.

CONCLUSION

Common measures bias appears to be partially mitigated by the interaction between organizational design and basic training. Further, training impacts the importance weights placed on non-financial performance metrics by participants. This increase in weights did not transfer to differential ratings in the second round due to a focus on only financial metrics. Results from a second experiment indicate this reluctance to change behavior may be strong, even for non-business majors.

While the results of the two experiments provide results that indicate training and organizational design can partially mitigate the common measures bias, I can conclude that the common measures may be more stubborn than was previously realized. The finding that participants were willing to re-evaluate the importance of metrics after training, but then subsequently ignore their own importance ratings in favor of a focus on financial metrics is especially concerning. Because the subjects in experiment one were accounting students, their exposure in previous classes could have had a shareholder primacy focus where profit maximization is the key, and their toolsets favor individual accounting and finance models in isolation. Khurna (2007) notes:

> "Inside business schools, economists on finance faculties used principal-agent theory to recast the role of management. Instead of being responsible to multiple stakeholders for the long-term well-being of the corporation, managers were now said to be responsible only to shareholders, a group whose composition changed continually and that was focused entirely on short-term gains. . . The resulting corporate oligarchy had no role-defined obligation other than to self-interest."

A Technology, Entertainment, and Design (TED) talk by Tom Wujec (2010) about his experience with team-based experiments that involve assimilation of different types of information mentions, ". . . among the worst performing teams are recent graduates of business school . . . the reason is that business students are trained to find the single right plan . . . and then they execute on it…..the best teams are those that are recent graduates of kindergarten." The logic given is that kindergarten students are willing to test many different prototypes to find a solution. In future experiments, I plan to use a subject pool outside the business school (e.g. English or Philosophy majors) to test whether business students' previous training has made it more difficult to assimilate non-financial information into their decisions. Interestingly, while the participants in the second experiment didn't show an initial common measures bias, they exhibited similar behavior to participants in experiment one in ignoring their own re-evaluation of the importance of certain metrics when making decisions.

There are some limitations to this investigation. First, participants may not have fully understood the prompt on organizational design. The predictions on the effect of organizational design on performance ratings are complex. The prompt given may not have fully articulated the intentions of the prompt and how the prompt would lead to the differential treatment of the metrics in decision making. Second, my initial training prompt was a very short video on the BSC concept that all metrics have incremental value and information for decisions. While this training did cause a change in the importance ratings of categories of non-financial metrics, it does not constitute the full extent of traditional BSC training. Within a controlled experiment there were limited options to conduct lengthy BSC training exercise. The second experiment attempts to push this narrative further by increasing the length and depth of training; however

further investigation can isolate other aspects of traditional BSC training to identify differential effects those training aspects have on performance evaluations and the common measures bias.

APPENDICES

APPENDIX A:

Overall Scenario

Welcome to WCS Inc., a global leader in apparel. I hope you are adjusting to your new position as personnel supervisor. As part of this position, you are tasked with evaluating the performance of two division managers: John who manages the RadWear division and Steve who manages the WorkWear division.

WorkWear sources and sells business apparel to young professionals. WorkWear's management determined that its growth must take place through focusing on the high-end market. As such, they have implemented initiatives to provide a more customer-focused experience. WorkWear also determined that it must increase the number of designer brands offered to keep the attention and capture the clothing dollars of its young professional customers. In addition, WorkWear has initiated a personal shopper program where sales personnel interact with and determine a customer's style preferences to then provide monthly clothing options tailored to the customer's personal style.

RadWear sources and sells sports apparel to young adults. RadWear has determined that increasing sales should come through an increased online presence. This increased online presence allows RadWear to customize orders at the point of manufacture to include special orders such as monogrammed clothing. The focus on online sales allows RadWear to shift costs away from inventory storage to provide a better quality product to the customer while simultaneously increasing variety in the products (color and size options).

APPENDIX B:

Competitive Prompt

The overall strategy of WCS Inc. is one of competition where divisions are encouraged to directly compete to achieve the goals of the firm. To do this, WCS has implemented an optimization program where efficiencies within divisions are encouraged. The competitive environment allows successful divisions to rise to the top and provide value to WCS. Each division has its own strategy to achieve its goals.

APPENDIX C:

Cooperative Prompt

The overall strategy of WCS Inc. is one of cooperation where divisions are encouraged to work together to achieve the goals of the firm. To do this, WCS has implemented a program of cross-sales where customer lists are shared between departments and special coupons are sent to encourage customers to buy from additional divisions. Additionally, WCS has implemented monthly meetings where divisional managers share strategies for improving operations. Each division has its own strategy to achieve its goals.

APPENDIX D:

Decision Material

TABLE 1: PERFORMANCE OF WORKWEAR DIVISION

Measure	Target	Actual
Financial:		
1. Return on sales	24%	25.9%
2. New store sales	30%	30.6%
3. Sales growth	35%	37.8%
4. Return relative to retail space	$80	$81.60
Customer-Related:		
1. Personal shopper program rating	85%	86.7%
2. Repeat sales	30%	32.4%
3. Returns by customers as a percentage of sales	12%	11.8%
4. Customer satisfaction rating	92%	99.4%
Internal Business Processes:		
1. Returns to suppliers	6%	5.5%
2. Average major brand names per store	32	32.6
3. Average markdowns	16%	14.7%
4. Sales from new market leaders	25%	25.5%
Learning and Growth:		
1. Average tenure of sales personnel	3	3.1
2. Hour of employee training per employee	15	16.2
3. Average customers per sales personnel	25	24.5
4. Employee suggestions per employee	3.3	3.6

Indicate your performance evaluation by writing a number on the line below.

0 50 100

|- - - | - - - | - - - | - - - | - - - | - - - | - - - | - - - | - - - | - - - |

Very Poor Poor Average Good Excellent

Rating _____

TABLE 2: PERFORMANCE OF RADWEAR DIVISION

Measure	Target	Actual
Financial:		
1. Return on sales	24%	24.5%
2. Revenues per online sales visit	$50	$51.00
3. Sales growth	35%	35.7%
4. Percentage of online sales	80%	81.6%
Customer-Related:		
1. Customer rating of online site	85%	86.7%
2. Repeat sales	30%	30.6%
3. Website views without purchases	20%	19.6%
4. Customer satisfaction rating	92%	93.8%
Internal Business Processes:		
1. Returns to suppliers	6%	5.9%
2. Orders filled within one week	90%	91.8%
3. Average markdowns	16%	15.7%
4. Online orders filled with errors	5%	4.9%
Learning and Growth:		
1. Percent of employees with programming skills	60%	61.2%
2. Hour of employee training per employee	15	15.3
3. Stores with online kiosks	75%	76.5%
4. Employee suggestions per employee	3.3	3.4

Indicate your performance evaluation by writing a number on the line below.

0 50 100

|- - - | - - - | - - - | - - - | - - - | - - - | - - - | - - - | - - - | - - - |

Very Poor Poor Average Good Excellent

Rating _____

Allocation Decision

As a corporate supervisor, you have discretion over capital investment decisions. The company has $100,000 this period with which to make capital investments. How will you divide the $100,000 between the two divisions?

RadWear Division $_____

WorkWear Division $_____

Please circle which manager did better (circle only one).

RadWear Manager WorkWear Manager

APPENDIX E:

Post-Round Questionnaire

Post-Round Questionnaire

Please indicate the importance you placed on each performance measure in making your performance evaluation decisions, where 1 equals low importance and 7 equals high importance. Please place an x or circle to indicate an importance for each metric below for the WorkWear Division.

TABLE 3: WORKWEAR DIVISION POST-ROUND QUESTIONNAIRE

Measure	Importance						
Financial:							
1. Return on sales	1	2	3	4	5	6	7
2. New store sales	1	2	3	4	5	6	7
3. Sales growth	1	2	3	4	5	6	7
4. Return relative to retail space	1	2	3	4	5	6	7
Customer-Related:							
1. Personal shopper program rating	1	2	3	4	5	6	7
2. Repeat sales	1	2	3	4	5	6	7
3. Returns by customers as a percentage of sales	1	2	3	4	5	6	7
4. Customer satisfaction rating	1	2	3	4	5	6	7
Internal Business Processes:							
1. Returns to suppliers	1	2	3	4	5	6	7
2. Average major brand names per store	1	2	3	4	5	6	7
3. Average markdowns	1	2	3	4	5	6	7
4. Sales from new market leaders	1	2	3	4	5	6	7
Learning and Growth:							
1. Average tenure of sales personnel	1	2	3	4	5	6	7
2. Hour of employee training per employee	1	2	3	4	5	6	7
3. Average customers per sales personnel	1	2	3	4	5	6	7
4. Employee suggestions per employee	1	2	3	4	5	6	7

Post-Round Questionnaire

Please indicate the importance you placed on each performance measure in making your performance evaluation decisions, where 1 equals low importance and 7 equals high importance. Please place an x or circle to indicate an importance for each metric below for the RadWear Division.

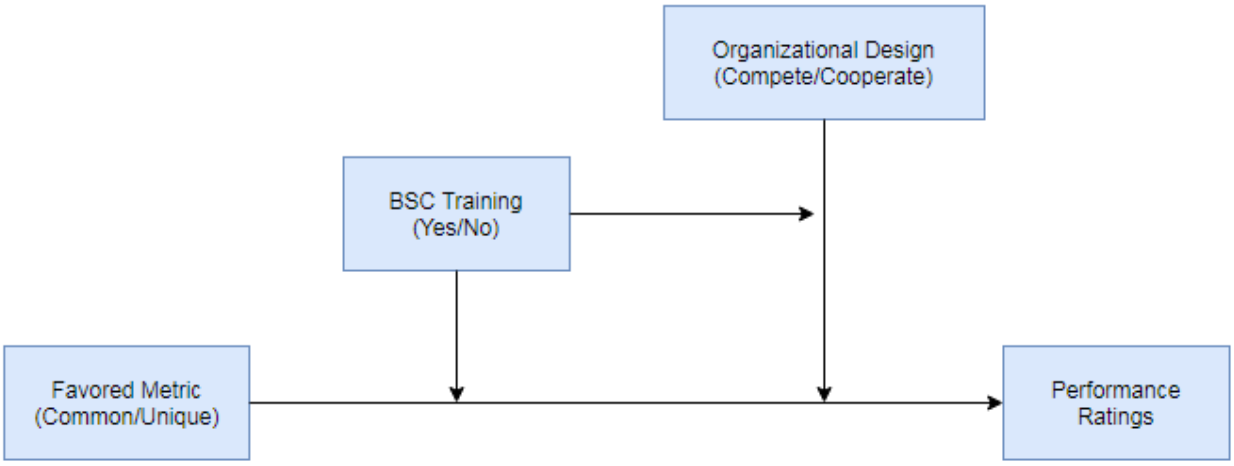
TABLE 4: RADWEAR DIVISION POST-ROUND QUESTIONNAIRE

Measure	Importance						
Financial:							
1. Return on sales	1	2	3	4	5	6	7
2. Revenues per online sales visit	1	2	3	4	5	6	7
3. Sales growth	1	2	3	4	5	6	7
4. Percentage of online sales	1	2	3	4	5	6	7
Customer-Related:							
1. Customer rating of online site	1	2	3	4	5	6	7
2. Repeat sales	1	2	3	4	5	6	7
3. Website views without purchases	1	2	3	4	5	6	7
4. Customer satisfaction rating	1	2	3	4	5	6	7
Internal Business Processes:							
1. Returns to suppliers	1	2	3	4	5	6	7
2. Orders filled within one week	1	2	3	4	5	6	7
3. Average markdowns	1	2	3	4	5	6	7
4. Online orders filled with errors	1	2	3	4	5	6	7
Learning and Growth:							
1. Percent of employees with programming skills	1	2	3	4	5	6	7
2. Hour of employee training per employee	1	2	3	4	5	6	7
3. Stores with online kiosks	1	2	3	4	5	6	7
4. Employee suggestions per employee	1	2	3	4	5	6	7

APPENDIX F:

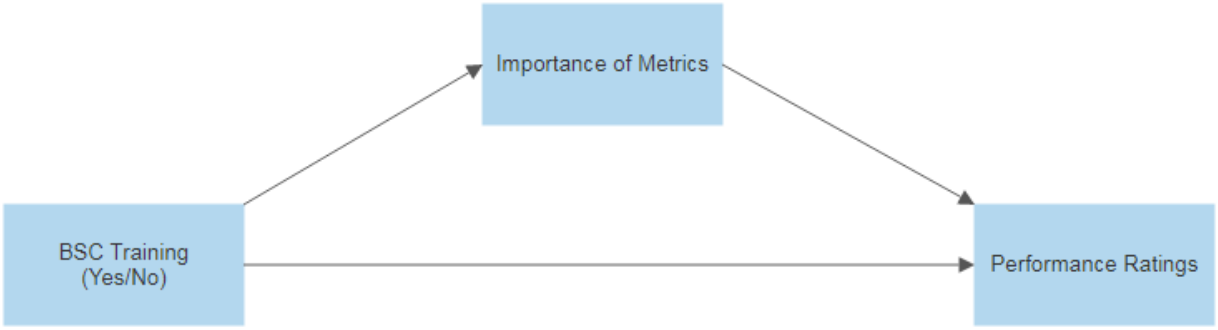
Figures

FIGURE 1: MODEL OF THE EFFECTS OF ORGANIZATIONAL DESIGN, PERFORMANCE METRICS, AND BSC TRAINING ON PERFORMANCE EVALUATIONS



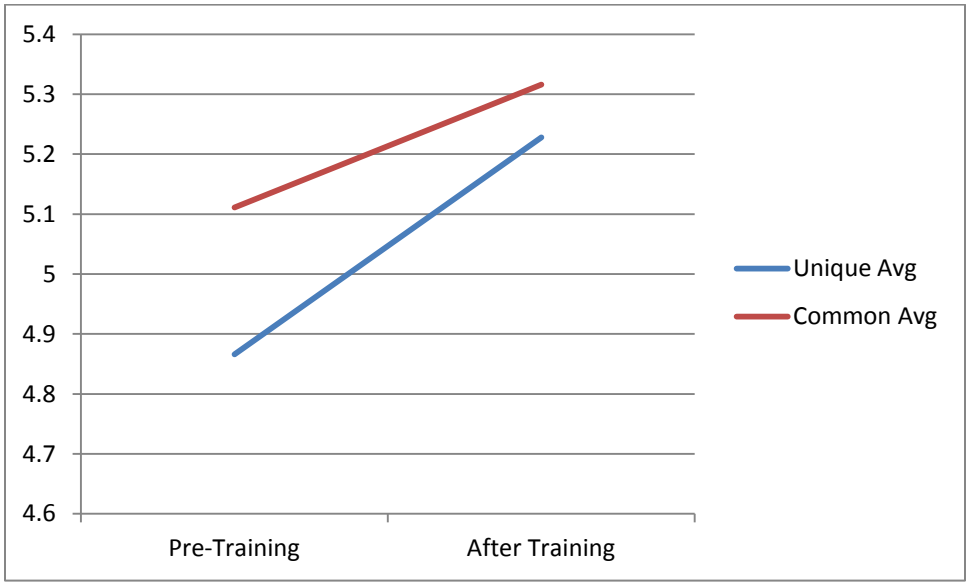
This model shows my predictions of the effects of organizational design (captured as whether divisions are designed to compete or cooperate), favored metrics (whether performance above target is higher for WorkWear on common or unique metrics), and BSC training (present or absent) on performance evaluations. Results for the analysis of the full model are presented in Table 7.

FIGURE 2: MEDIATED MODEL OF THE EFFECT OF TRAINING ON PERFORMANCE EVALUATIONS



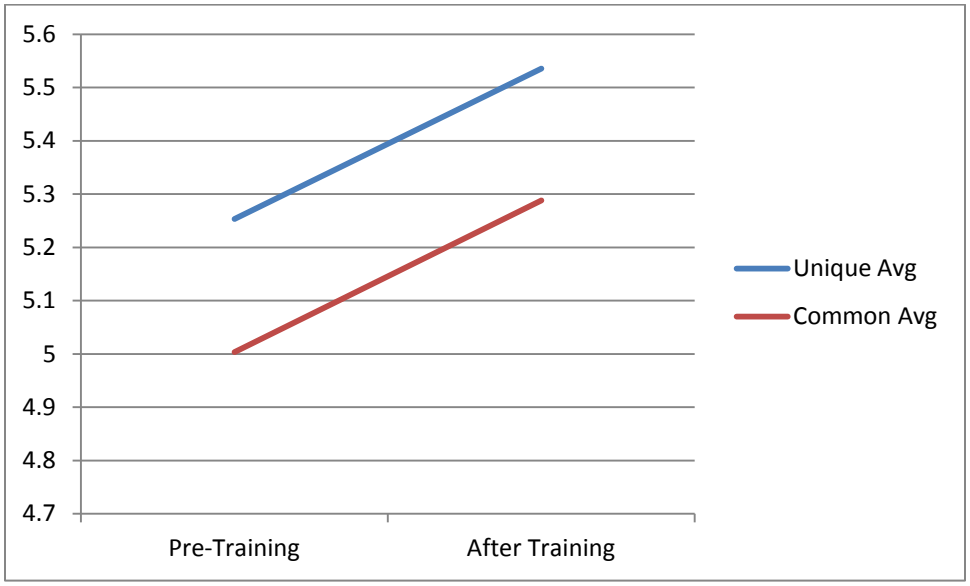
This model shows the general mediated effect of training on performance evaluations through the ratings on the importance of the metrics. Participants rated all 16 metrics for each division. I analyze this model from two perspectives. First, I look at the average ratings for each of the two division's BSC categories. This gives 8 mediation variables to include in the model (Financial, Customer Perspective, Internal Business, and Innovation for both WorkWear and RadWear). Second, I look at the average ratings for unique and common metrics for each of the two division. This gives 4 mediation variables to include in the model (Common WorkWear, Unique WorkWear, Common RadWear, and Unique RadWear). Results of the analysis of this model are presented in Tables 9 and 10.

FIGURE 3: EXPERIMENT 1 – WORKWEAR IMPORTANCE OF METRIC RATINGS



The difference in slopes between the common and unique metrics for WorkWear is significant at the 5% level.

FIGURE 4: EXPERIMENT 1 – RADWEAR IMPORTANCE OF METRIC RATINGS



The difference in slopes between the common and unique metrics for RadWear is insignificant.

APPENDIX G:

Tables

TABLE 5: COMMON AND UNIQUE PERFORMANCE MEASURES FOR WORKWEAR AND RADWEAR BALANCED SCORECARDS

Measure	Type
Financial:	
Return on sales	C
Sales growth	C
New store sales	U-Work
Return relative to retail space	U-Work
Percentage of online sales	U-Rad
Revenue per online sales visit	U-Rad
Customer-Related:	
Customer satisfaction rating	C
Repeat sales	C
Personal shopper program rating	U-Work
Returns by customers as a percentage of sales	U-Work
Website views without purchases	U-Rad
Customer rating of online site	U-Rad
Internal Business Processes:	
Returns to suppliers	C
Average markdowns	C
Average major brand names per store	U-Work
Sales from new market leaders	U-Work
Online orders filled with errors	U-Rad
Orders filled within one week	U-Rad
Learning and Growth:	
Employee suggestions per employee	C
Hour of employee training per employee	C
Average customers per sales personnel	U-Work
Average tenure of sales personnel	U-Work
Percent of employees with programming skills	U-Rad
Stores with online kiosks	U-Rad

C denotes common metrics present on both division's Balanced Scorecards. U-Work denotes metrics present only on WorkWear's Balanced Scorecard. U-Rad denotes metrics present only on RadWear's Balanced Scorecard.

TABLE 6: EXPERIMENT 1 – EXPERIMENTAL RESULTS FOR MANAGERS' PERFORMANCE EVALUATIONS IN THE FIRST ROUND

Panel A: Results of 2 x 2 ANOVA of Evaluations of the Performance of WorkWear and RadWear Divisions' Managers

Variable	df	SS	MS	F	p
Metric	1	231.72	231.72	4.27	0.041
Compete	1	21.29	21.29	0.39	0.533
Metric x Compete	1	12.92	12.92	0.24	0.627
Error	104	5650.19	54.33		

Panel B: Comparison of Means for WorkWear and RadWear Divisions' Managers

Competitive Organizational Design Division	Favor Common Metrics	Favor Unique Metrics
WorkWear	85.85	81.62
	(8.37)	(10.32)
Rad Wear	80.59	81.85
	(11.71)	(9.45)
Difference:WorkWear - RadWear[a]	5.26*	-0.23

Cooperative Organizational Design Division	Favor Common Metrics	Favor Unique Metrics
WorkWear	84.46	80.74
	(10.42)	(8.54)
RadWear	79.32	79.70
	(10.64)	(10.19)
Difference: WorkWear - RadWear[b]	5.14*	1.04

Performance Evaluations were made on a 101-point scale. Panel B values are means (standard deviations). Favor Common (Unique) Metrics gives the conditions where performance above targets for WorkWear division were higher for Common (Unique) Metrics. Significance levels indicate where p-values are: * < 10%, ** < 5%, *** <1%.

a: The difference for the Favor Common Metrics condition is significant at the 10% level, while the difference for the Favor Unique Metrics condition is insignificant.

b: The difference for the Favor Common Metrics condition is significant at the 10% level, while the difference for the Favor Unique Metrics condition is insignificant.

TABLE 7: EXPERIMENT 1 – EXPERIMENTAL RESULTS FOR MANAGERS' PERFORMANCE EVALUATIONS IN THE SECOND ROUND

Panel A: Results of 2 x 2 ANOVA of Evaluations of the Performance of WorkWear and RadWear Divisions' Managers

Variable	df	SS	MS	F	p
Metric	1	21.42	21.42	0.20	0.655
Compete	1	101.91	101.91	0.96	0.330
Training	1	55.45	55.45	0.52	0.472
Compete x Metric	1	112.44	112.44	1.06	0.306
Compete x Training	1	1.39	1.39	0.01	0.909
Metric x Training	1	73.17	73.17	0.69	0.409
Compete x Metric x Training	1	35.31	35.31	0.33	0.566
Error	100	10631.81	106.32		

Panel B: Comparison of Means for WorkWear and RadWear Divisions' Managers

Competitive Organizational Design Division	Favor Common Metrics	Favor Unique Metrics
WorkWear	86.48	85.69
	(8.56)	(7.95)
Rad Wear	79.46	81.69
	(12.12)	(7.99)
Difference:WorkWear - RadWear[a]	7.02**	4.00*

Cooperative Organizational Design Division	Favor Common Metrics	Favor Unique Metrics
WorkWear	86.21	81.41
	(10.95)	(9.22)
RadWear	79.86	79.70
	(11.84)	(9.71)
Difference: WorkWear - RadWear[b]	6.35**	1.71

Performance Evaluations were made on a 101-point scale. Panel B values are means (standard deviations). Favor Common (Unique) Metrics gives the conditions where performance above targets for WorkWear division were higher for Common (Unique) Metrics. Significance levels indicate where p-values are: * < 10%, ** < 5%, *** <1%.
a: The difference for the Favor Common Metrics condition is significant at the 5% level, while the difference for the Favor Unique Metrics condition is significant at the 10% level.
b: The difference for the Favor Common Metrics condition is significant at the 5% level, while the difference for the Favor Unique Metrics condition is insignificant.

TABLE 8: EXPERIMENT 1 – METRIC IMPORTANCE RATINGS

Financial Metrics	First Round Importance			Second Round Importance			R2 – R1	
Division	Common	Unique	Difference	Common	Unique	Difference	Common	Unique
WorkWear	6.097	5.093	1.004***	6.134	5.310	0.824***	0.037	0.217*
Rad Wear	6.125	5.958	0.167	6.204	6.069	0.135	0.079	0.111
Difference	-0.028	-0.865***		-0.070	-0.759***			
Customer Metrics								
Division	Common	Unique	Difference	Common	Unique	Difference	Common	Unique
WorkWear	6.116	5.333	0.783***	6.185	5.671	0.514***	0.069	0.338**
RadWear	5.889	5.505	0.384***	6.088	5.692	0.396***	0.199*	0.187
Difference	0.227**	-0.172		0.097	-0.021			
Internal Business Metrics								
Division	Common	Unique	Difference	Common	Unique	Difference	Common	Unique
WorkWear	4.130	4.523	-0.393**	4.412	4.921	-0.509***	0.282*	0.398**
RadWear	4.037	5.102	-1.065***	4.417	5.468	-1.051***	0.380**	0.366**
Difference	0.093	-0.579***		-0.005	-0.547***			
Growth Metrics								
Division	Common	Unique	Difference	Common	Unique	Difference	Common	Unique
WorkWear	4.102	4.514	-0.412**	4.532	5.009	-0.477**	0.430**	0.495***
RadWear	3.958	4.449	-0.491***	4.444	4.912	-0.468**	0.486***	0.463**
Difference	0.144	0.065		0.088	0.097			

Numbers represented are mean importance ratings values for common and unique metrics in each of the four Balanced Scorecard categories. Significance levels indicate where p-values are: * < 10%, ** < 5%, *** <1%. The third set of columns (R2-R1) represents the change in rating between round 1 and round 2.

TABLE 9: EXPERIMENT 1 – DIRECT AND INDIRECT EFFECT OF TRAINING ON RATING CHANGES THROUGH CHANGES IN THE AVERAGE IMPORTANCE OF BSC CATEGORIES

IV Effect on Mediators:	Coefficient	Standard Error
Training -> Δ WorkWear Financial Metric	-0.058	(0.115)
Training -> Δ WorkWear Customer Metric	0.282**	(0.119)
Training -> Δ WorkWear Internal Business Metric	0.317**	(0.160)
Training -> Δ WorkWear Innovation Metric	0.375*	(0.212)
Training -> Δ RadWear Financial Metric	0.054	(0.103)
Training -> Δ RadWear Customer Metric	0.081	(0.127)
Training -> Δ RadWear Internal Business Metric	0.120	(0.182)
Training -> Δ RadWear Innovation Metric	0.445**	(0.188)
Total Effect on DV (Δ Rating Difference)		
Δ WorkWear Financial Metric	3.513*	(1.890)
Δ WorkWear Customer Metric	1.999	(1.658)
Δ WorkWear Internal Business Metric	1.513	(1.515)
Δ WorkWear Innovation Metric	-1.320	(1.257)
Δ RadWear Financial Metric	-0.813	(2.210)
Δ RadWear Customer Metric	-1.946	(1.616)
Δ RadWear Internal Business Metric	-0.941	(1.386)
Δ RadWear Innovation Metric	-0.758	(1.313)
Training	0.888	(1.988)

The above table shows the results of analyzing the model where training is mediated through the importance ratings of the BSC categories of metrics for both divisions as shown in Figure 2. The individual effects of training on each of the mediating variables are presented as "IV Effect on Moderators". The direct effect of training as well as the effects of the mediating variables on the DV of change in performance ratings are presented in the "Total Effects" portion. The mediating variables measure the change in average rating for each category of metrics between the first and second round. Training is a dichotomous variable on whether BSC training was present or absent. The DV is measured as the change in rating difference between the two divisions from the first to the second round as shown in the following equation:
Δ Rating Difference = (Round 2 WorkWear Rating – Round 2 RadWear Rating) - (Round 1 WorkWear Rating – Round 1 RadWear Rating)
Significance levels indicate where p-values are: * < 10%, ** < 5%, *** <1%.

TABLE 10: EXPERIMENT 1 – DIRECT AND INDIRECT EFFECT OF TRAINING ON RATING CHANGES THROUGH CHANGES IN IMPORTANCE OF UNIQUE AND COMMON METRICS

IV Effect on Mediators:	Coefficient	Standard Error
Training -> Δ WorkWear Common Metrics	0.253**	(0.104)
Training -> Δ WorkWear Unique Metrics	0.205*	(0.106)
Training -> Δ RadWear Common Metrics	0.268***	(0.101)
Training -> Δ RadWear Unique Metrics	0.082	(0.124)
Total Effect on DV (Δ Rating Difference)		
Δ WorkWear Common Metrics	-0.501	(2.951)
Δ WorkWear Unique Metrics	2.766	(2.280)
Δ RadWear Common Metrics	-2.992	(2.823)
Δ RadWear Unique Metrics	-2.570	(1.777)
Training	1.153	(2.004)

The above table shows the results of analyzing the model where training is mediated through the importance ratings of the type of metrics for both divisions as shown in Figure 2. The individual effects of training on each of the mediating variables are presented as "IV Effect on Moderators". The direct effect of training as well as the effects of the mediating variables on the DV of change in performance ratings are presented in the "Total Effects" portion. The mediating variables measure the change in average rating for common and unique metrics for each of the two divisions from round 1 to round 2. Training is a dichotomous variable on whether BSC training was present or absent. The DV is measured as the change in rating difference between the two divisions from the first to the second round as shown in the following equation:
Δ Rating Difference = (Round 2 WorkWear Rating – Round 2 RadWear Rating) - (Round 1 WorkWear Rating – Round 1 RadWear Rating)
Significance levels indicate where p-values are: * < 10%, ** < 5%, *** <1%.

TABLE 11: EXPERIMENT 2 – EXPERIMENTAL RESULTS FOR MANAGERS' PERFORMANCE EVALUATIONS IN THE FIRST ROUND

Panel A: Results of 2 x 2 ANOVA of Evaluations of the Performance of WorkWear and RadWear Divisions' Managers

Variable	df	SS	MS	F	p
Metric	1	23.113	23.113	0.30	0.583
Error	78	5928.375	76.005		

Panel B: Comparison of Means for WorkWear and RadWear Divisions' Managers

Division	Favor Common Metrics	Favor Unique Metrics
WorkWear	82.83	82.50
	(12.01)	(12.25)
Rad Wear	76.43	77.18
	(15.17)	(12.98)
Difference:WorkWear - RadWear[a]	6.40**	5.32*

Performance Evaluations were made on a 101-point scale. Panel B values are means (standard deviations). Favor Common (Unique) Metrics gives the condition where performance above target for WorkWear division was higher than for RadWear division, keeping Unique (Common) metric performance constant. Significance levels indicate where p-values are: * < 10%, ** < 5%, *** <1%.

a: The differences for the Favor Common Metrics and for the Favor Unique Metrics condition are both significant.

TABLE 12: EXPERIMENT 2 – EXPERIMENTAL RESULTS FOR MANAGERS' PERFORMANCE EVALUATIONS IN THE SECOND ROUND

Panel A: Results of 2 x 2 ANOVA of Evaluations of the Performance of WorkWear and RadWear Divisions' Managers

Variable	df	SS	MS	F	p
Metric	1	18.421	18.421	0.26	0.610
Training	1	5.625	5.625	0.08	0.778
Metric x Training	1	0.017	0.017	0.00	0.988
Error	76	5326.685	70.088		

Panel B: Comparison of Means for WorkWear and RadWear Divisions' Managers

Division	Favor Common Metrics	Favor Unique Metrics
WorkWear	82.60	82.38
	(13.81)	(11.98)
Rad Wear	74.86	77.05
	(15.30)	(13.69)
Difference:WorkWear - RadWear[a]	7.74**	5.33*

Performance Evaluations were made on a 101-point scale. Panel B values are means (standard deviations). Favor Common (Unique) Metrics gives the conditions where performance above targets for WorkWear division were higher for Common (Unique) Metrics. Significance levels indicate where p-values are: * < 10%, ** < 5%, *** <1%.
a: The differences for the Favor Common Metrics and for the Favor Unique Metrics condition are both significant.

TABLE 13: EXPERIMENT 2 – DIRECT AND INDIRECT EFFECT OF TRAINING ON RATING CHANGES THROUGH CHANGES IN THE AVERAGE IMPORTANCE OF BSC CATEGORIES

IV Effect on Mediators:	Coefficient	Standard Error
Training -> Δ WorkWear Financial Metric	-0.396**	(0.160)
Training -> Δ WorkWear Customer Metric	-0.161	(0.148)
Training -> Δ WorkWear Internal Business Metric	-0.047	(0.174)
Training -> Δ WorkWear Innovation Metric	0.316*	(0.185)
Training -> Δ RadWear Financial Metric	-0.343**	(0.148)
Training -> Δ RadWear Customer Metric	-0.126	(0.161)
Training -> Δ RadWear Internal Business Metric	0.029	(0.158)
Training -> Δ RadWear Innovation Metric	0.431***	(0.156)
Total Effect on DV (Δ Rating Difference)		
Δ WorkWear Financial Metric	2.436*	(1.364)
Δ WorkWear Customer Metric	2.175	(1.620)
Δ WorkWear Internal Business Metric	0.657	(1.288)
Δ WorkWear Innovation Metric	-2.574*	(1.436)
Δ RadWear Financial Metric	-0.972	(1.496)
Δ RadWear Customer Metric	0.784	(1.412)
Δ RadWear Internal Business Metric	-0.832	(1.453)
Δ RadWear Innovation Metric	-0.315	(1.553)
Training	2.797	(1.939)

The above table shows the results of analyzing the model where training is mediated through the importance ratings of the BSC categories of metrics for both divisions as shown in Figure 2. The individual effects of training on each of the mediating variables are presented as "IV Effect on Moderators". The direct effect of training as well as the effects of the mediating variables on the DV of change in performance ratings is presented in the "Total Effects" portion. The mediating variables measure the change in average rating for each category of metrics between the first and second round. Training is a dichotomous variable on whether BSC training was present or absent. The DV is measured as the change in rating difference between the two divisions from the first to the second round as shown in the following equation:
Δ Rating Difference = (Round 2 WorkWear Rating – Round 2 RadWear Rating) - (Round 1 WorkWear Rating – Round 1 RadWear Rating)
Significance levels indicate where p-values are: * < 10%, ** < 5%, *** <1%.

TABLE 14: EXPERIMENT 2 – DIRECT AND INDIRECT EFFECT OF TRAINING ON RATING CHANGES THROUGH CHANGES IN IMPORTANCE OF UNIQUE AND COMMON METRICS

IV Effect on Mediators:	Coefficient	Standard Error
Training -> Δ WorkWear Common Metrics	-0.069	(0.126)
Training -> Δ WorkWear Unique Metrics	-0.075	(0.130)
Training -> Δ RadWear Common Metrics	0.047	(0.119)
Training -> Δ RadWear Unique Metrics	-0.051	(0.105)
Total Effect on DV (Δ Rating Difference)		
Δ WorkWear Common Metrics	1.451	(2.046)
Δ WorkWear Unique Metrics	-0.127	(1.786)
Δ RadWear Common Metrics	-0.235	(2.123)
Δ RadWear Unique Metrics	-1.554	(2.122)
Training	0.734	(1.837)

The above table shows the results of analyzing the model where training is mediated through the importance ratings of the type of metrics for both divisions as shown in Figure 2. The individual effects of training on each of the mediating variables are presented as "IV Effect on Moderators". The direct effect of training as well as the effects of the mediating variables on the DV of change in performance ratings is presented in the "Total Effects" portion. The mediating variables measure the change in average rating for common and unique metrics for each of the two divisions from round 1 to round 2. Training is a dichotomous variable on whether BSC training was present or absent. The DV is measured as the change in rating difference between the two divisions from the first to the second round as shown in the following equation:
Δ Rating Difference = (Round 2 WorkWear Rating – Round 2 RadWear Rating) - (Round 1 WorkWear Rating – Round 1 RadWear Rating)
Significance levels indicate where p-values are: * < 10%, ** < 5%, *** <1%.

TABLE 15: EXPERIMENT 2 – METRIC IMPORTANCE RATINGS

Financial Metrics	First Round Importance			Second Round Importance			R2 – R1	
Division	Common	Unique	Difference	Common	Unique	Difference	Common	Unique
WorkWear	5.900	5.188	0.712***	5.869	5.250	0.619***	-0.031	0.062
Rad Wear	5.788	5.550	0.238	5.713	5.713	0.000	-0.075	0.163
Difference	0.112	-0.362**		0.156	-0.463***			
Customer Metrics								
Division	Common	Unique	Difference	Common	Unique	Difference	Common	Unique
WorkWear	5.788	5.050	0.738***	5.825	5.350	0.475***	0.037	0.300*
RadWear	5.625	5.031	0.594***	5.731	5.306	0.425***	0.106	0.275*
Difference	0.163	0.019		0.094	0.044			
Internal Business Metrics								
Division	Common	Unique	Difference	Common	Unique	Difference	Common	Unique
WorkWear	4.681	4.618	0.063	4.600	4.756	-0.156	-0.081	0.138
RadWear	4.700	5.013	-0.313	4.700	5.175	-0.475***	0.000	0.162
Difference	-0.019	-0.395*		-0.100	-0.419**			
Growth Metrics								
Division	Common	Unique	Difference	Common	Unique	Difference	Common	Unique
WorkWear	4.513	4.731	-0.218	4.844	4.969	-0.125	0.331	0.238
RadWear	4.481	4.481	0.000	4.719	4.525	0.194	0.238	0.044
Difference	0.032	0.250		0.125	0.444			

Numbers represented are mean importance ratings values for common and unique metrics in each of the four Balanced Scorecard categories. Significance levels indicate where p-values are: * < 10%, ** < 5%, *** <1%. The third set of columns (R2-R1) represents the change in rating between round 1 and round 2.

REFERENCES

REFERENCES

Baker, George. 2002. "Distortion and Risk in Optimal Incentive Contracts." *The Journal of Human Resources* 37 (4): 728–51.

Banker, Rajiv, Chang, Hsihui, and Pizzini, Mina. 2004. "The Balanced Scorecard: Judgmental Effects of Performance Measures Linked to Strategy." *The Accounting Review* 79 (1): 1–23.

Bobb, Brooke. 2019 (July 30). "Amazon Fashion Launches a Personal Styling Service – Will Customers Actually Use it?" *Vogue.com* Retrieved from https://www.vogue.com/article/amazon-prime-personal-shopper?verso=true.

Calabro, Lori. 2001 (February 1). "On Balance." *CFO.com* Retrieved from https://www.cfo.com/strategy/2001/02/on-balance/.

Cardinaels, Eddy, and van Veen-Dirks, Paula. 2010. "Financial versus Non-Financial Information: The Impact of Information Organization and Presentation in a Balanced Scorecard." *Accounting, Organizations and Society* 35 (6): 565–78.

Chenhall, Robert, and Morris, Deigan. 1986. "The Impact of Structure, Environment, and Interdependence on the Perceived Usefulness of Management Accounting Systems." *The Accounting Review* 61 (1): 16–35.

Crabtree, Aaron, and DeBusk, Gerald. 2008. "The effects of adopting the balanced scorecard on shareholder returns." *Advances in Accounting* 24 (1), 8-15.

Croskerry, Pat, Singhal, Geeta, and Mamede, Silvia. 2013. "Cognitive Debiasing 1: Origins of Bias and Theory of Debiasing." *BMJ Quality & Safety* 22 (Suppl 2): ii58–64.

Dewey, John. 1938. *Experience and Education*. New York; Collier.

Dikolli, Shane, and Sedatole, Karen. 2007. "Improvements in the Information Content of Nonfinancial Forward-Looking Performance Measures: A Taxonomy and Empirical Application." *Journal of Management Accounting Research* 19 (1): 71–104.

Dilla, William and Steinbart, Paul. 2005. "Relative Weighting of Common and Unique Balanced Scorecard Measures by Knowledgeable Decision Makers." *Behavioral Research in Accounting* 17 (1): 43–53.

Epstein, Marc, and Roy, Marie-Josee. 2005. "Evaluating and monitoring CEO performance: evidence from US compensation committee reports." *Corporate Governance* 5 (4), 75-87.

Frederiks, Elisha, Stenner, Karen, and Hobman, Elizabeth. 2015. "Household Energy Use: Applying Behavioural Economics to Understand Consumer Decision-Making and Behaviour." *Renewable and Sustainable Energy Reviews* 41 (January): 1385–94.

Frontczak, Nancy T. 1998. "A Paradigm for the Selection, Use and Development of Experiential Learning Activities in Marketing Education." *Marketing Education Review* 8 (3): 25–33.

Govindarajan, Vijay and Gupta, Anil. 1985. "Linking Control Systems to Business Unit Strategy: Impact on Performance." *Accounting, Organizations and Society* 10 (1): 51–66.

Gupta, Anil, and Govindarajan, Vijay. 1986. "Resource Sharing among SBUs: Strategic Antecedents and Administrative Implications." *The Academy of Management Journal* 29 (4): 695–714.

Hibbets, Aleecia, Roberts, Michael, and Albright, Thomas. 2006. "Common-Measures Bias in the Balanced Scorecard: Cognitive Effort and General Problem-Solving Ability." Unpublished SSRN paper id=921473.

Hill, Charles, Hitt, Michael, and Hoskisson, Robert. 1992. "Cooperative Versus Competitive Structures in Related and Unrelated Diversified Firms." *Organization Science* 3 (4): 501–21.

Hoque, Zahirul. 2004. "A Contingency Model of the Association between Strategy, Environmental Uncertainty and Performance Measurement: Impact on Organizational Performance." *International Business Review* 13 (4): 485–502.

Humphreys, Kerry and Trotman, Ken. 2011. "The Balanced Scorecard: The Effect of Strategy Information on Performance Evaluation Judgments." *Journal of Management Accounting Research* 23 (1): 81–98.

Hunt, Jasper. 1995. "Dewey's Philosophical Method and Its Influence on His Philosophy of Education," In *The Theory of Experiential Education* Warren, Karen, Sakofs, Mitchell, and Hunt, Jasper (eds.). 23-32. Kendall-Hunt Publishing Company.

Ittner, Christopher and Daid F. Larcker, David. 1998. "Are Nonfinancial Measures Leading Indicators of Financial Performance? An Analysis of Customer Satisfaction." *Journal of Accounting Research* 36: 1–35.

Jensen, Michael and Meckling, William. 1992. "Specific and General Knowledge and Organizational Structure." SSRN Scholarly Paper ID 6658. Rochester, NY: Social Science Research Network.

Jones, Charisse. 2018 (May 31). "Walmart shoppers can summon a personal shopper with a text – and for $50 a month." *USAToday.com*. Retrieved from https://www.usatoday.com/story/money/2018/05/31/walmart-personal-shoppers-amazon-rent-runway-text/661170002/

Kaplan, Robert and Norton, David. 1992. “The Balanced Scorecard--Measures That Drive Performance.” *Harvard Business Review* 70 (1): 71–79.

Kaplan, Robert and Norton, David. 1996. “Linking the Balanced Scorecard to Strategy.” *California Management Review; Berkeley* 39 (1): 53–79.

Kaplan, Steven and Wisner, Priscilla. 2009. “The Judgmental Effects of Management Communications and a Fifth Balanced Scorecard Category on Performance Evaluation.” *Behavioral Research in Accounting* 21 (2): 37–56.

Khurana, Rakesh. 2010. *From higher aims to hired hands: The social transformation of American business schools and the unfulfilled promise of management as a profession.* Princeton University Press.

Krumwiede, Kip, Swain, Monte, Thornock, Todd, and Eggett, Dennis. 2013. “The Effects of Task Outcome Feedback and Broad Domain Evaluation Experience on the Use of Unique Scorecard Measures.” *Advances in Accounting* 29 (2): 205–17.

Libby, Theresa, Salterio, Steven, and Webb, Alan. 2004. “The Balanced Scorecard: The Effects of Assurance and Process Accountability on Managerial Judgment.” *The Accounting Review* 79 (4): 1075–94.

Lipe, Marlys and Salterio, Steven. 2000. “The Balanced Scorecard: Judgmental Effects of Common and Unique Performance Measures.” *The Accounting Review* 75 (3): 283–98.

Lipe, Marlys and Salterio, Steven. 2002. “A Note on the Judgmental Effects of the Balanced Scorecard’s Information Organization.” *Accounting, Organizations and Society* 27 (6): 531–40.

Markman, Arthur and Medin, Douglas. 1995. “Similarity and Alignment in Choice.” *Organizational Behavior and Human Decision Processes* 63 (2): 117–30.

Miller, Craig, Nentl, Nancy, and Zietlow, Ruth. 2014. “About Simulations and Bloom’s Learning Taxonomy.” *Developments in Business Simulation and Experiential Learning: Proceedings of the Annual ABSEL Conference 37.*

Mosier, Kathleen, Skitka, Linda, Heers, Susan, and Burdick, Mark. 1998. “Automation Bias: Decision Making and Performance in High-Tech Cockpits.” *The International Journal of Aviation Psychology* 8 (1): 47–63.

Payne, John. 1976. “Task Complexity and Contingent Processing in Decision Making: An Information Search and Protocol Analysis.” *Organizational Behavior and Human Performance* 16 (2): 366–87.

Pitts, Robert. 1977. “Strategies And Structures For Diversification.” *Academy of Management Journal* 20 (2): 197–208.

Roberts, Michael, Albright, Thomas, and Hibbets, Aleecia. 2004. “Debiasing Balanced Scorecard Evaluations.” *Behavioral Research in Accounting* 16 (1): 75–88.

Schwenk, Charles. 1988. “The Cognitive Perspective on Strategic Decision Making.” *Journal of Management Studies* 25 (1): 41–55.

Slovic, Paul and MacPhillamy, Douglas. 1974. “Dimensional Commensurability and Cue Utilization in Comparative Judgment.” *Organizational Behavior and Human Performance* 11 (2): 172–94.

Stearns, James, and Crespy, Charles. 1995. “Learning Hierarchies and the Marketing Curriculum: A Proposal for a Second Course in Marketing.” *Journal of Marketing Education* 17 (2): 20–32.

Tennyson, Robert, and Merrill, David. 1971. “Hierarchical Models in the Development of a Theory of Instruction: A Comparison of Bloom, Gagne and Merrill.” *Educational Technology* 11 (9): 27–30.

Tversky, Amos and Kahneman, Daniel. 1974. Judgment under uncertainty: Heuristics and biases. S*cience, 185*(4157), 1124-1131.

Wiersma, Eelke. 2009. “For which purposes do managers use Balanced Scorecards?: An empirical study.” *Management Accounting Research* 20 (4), 239-251.

Wujec, Tom. 2010. “Build a Tower, Build a Team.” *TED2010 Talks*. Retrieved from: https://www.ted.com/talks/tom_wujec_build_a_tower?language=en.

Zimmerman, Jerold. 2011. *Accounting For Decision Making and Control Seventh Edition*. McGraw-Hill.