Common Measure Bias in the Balanced Scorecard:

an Experiment with Undergraduate Students

Kim H.M. Grevinga

MSc Business Administration
Financial Management

May 15, 2013

Title: Common Measure Bias in the Balanced Scorecard: an Experiment with Undergraduate Students

Student: *Kim H.M. Grevinga*
s1115561
k.h.m.grevinga@student.utwente.nl
Master MSc Business Administration – Financial Management
School of Management and Governance
University of Twente, Enschede

1st supervisor: *Dr. T. De Schryver*
School of Management and Governance

2nd supervisor: *Dr. Ir. J. Kraaijenbrink.*
School of Management and Governance

Period: November 2012 – May 2013

Preface

This thesis is the result of a study towards the influence of common measure bias on performance evaluation using a balanced scorecard. The study is conducted to complete the master Business Administration at the University of Twente.

An experiment with undergraduate students is conducted, to test whether common measure bias also hold in an experiment with undergraduate students. This study is also used to test whether undergraduate students could be used for further research in managerial accounting.

Finally, some words of acknowledgement are in place. First I would like to thank Tom de Schryver for giving me the opportunity to work on this project, but also for providing critical comments and useful tips during the research process. Also thanks to Jeroen Kraaijenbrink, for reading through the thesis, providing tips and taking care for the final review of the thesis. Furthermore, I would like to thank my parents and friends for their social support during my study.

Kim Grevinga
Hengelo, May 2013.

Abstract

In the early 1990s, the balanced scorecard (BSC) was introduced as a new tool for performance evaluation that complements the traditional financial performance measures with three other categories of performance measures (customer relations, internal business, and learning and growth). In this study, I explore how the cognitive limitations of decision makers' may prevent an organization to full benefit from the BSC. Observable characteristics of the BSC are examined (measures common to multiple divisions vs. measures unique to a particular division) that may limit decision makers in fully use the information provided by the BSC. This is tested using undergraduate students who were asked to compare two divisions of an organization on the basis of a BSC. These two divisions have different divisional strategies, and therefore the BSCs of these two divisions are different. Multiple BSCs with eight common measures and eight unique measures are used to compare the performance of the two divisional managers. In the experiment, the design of these BSCs is manipulated to test whether common measure bias is present. Previous research has found that decision makers tend to overweight the common measures of a BSC, which are used in the balanced scorecards of multiple divisions of an organization. Thus, these decision makers tend to ignore the unique measures of a BSC, which are the measures tailored to the strategy of that specific division. Overweighting the common measures and ignoring the unique measures instead, is called *common measure bias*.

Over the last decade, much research has been conducted on eliminating common measure bias in the BSC. In the literature review eight studies that have examined common measure bias in the BSC were explored. The comparison of these studies has led to the identification of five factors that could attenuate common measure bias in the BSC. This study contributes to prior literature of common measure bias because in the previous studies almost all experimental participants (most often M.B.A. students) have full-time work experience, which could influence the perception about performance measures.

The purpose of this study is to examine whether common measure bias as found in prior research holds in an experiment with undergraduate students to test whether they can be used for future research in managerial accounting. Based upon the literature review, three hypotheses are formulated. These three hypotheses are tested by means of an experiment with 207 undergraduate students with a mixed interest in management accounting. These students are trained in the use and design of the BSC, and could therefore be typified as knowledgeable decision makers with a theoretical understanding of the BSC. The experimental design of this study is based upon that of prior studies and asks students to evaluate the performance of two divisional managers from two separate divisions of the same organization. The balanced scorecards are manipulated to test whether performance on common and/or unique measures affect the overall judgment of decision makers. Furthermore, performance evaluation was linked to compensation decisions since it was aimed to test whether compensation decisions were affected by performance evaluations.

This study provides evidence that decision makers with a theoretical understanding of the BSC and relatively less full-time work experience, also incorporate unique measures in their performance evaluation. Significant interaction effects are found for as well the common performance measures

as the unique performance measures, which means that both the performance measures affect the evaluations of divisional managers. Furthermore, it is found that disaggregation of performance evaluation will attenuate common measure bias in the BSC. Also, it was found that compensation decisions (e.g. bonus allocations and promotion decisions) are affected by performance evaluation scores of divisional managers.

Thus, this study complements in several ways to the existing literature:

- No common measure bias was found for knowledgeable decision makers.
- Disaggregation of the BSC will lead to more attention for measures unique to one division.
- The BSC is particularly useful to link performance evaluation to compensation decisions.
- Undergraduate students were found to be useful in managerial accounting studies.

Practical implications are that when the decision makers are knowledgeable, the BSC is a useful tool for linking performance evaluation to compensation decisions. This thesis paves the way for researchers to use undergraduate students in future research on managerial accounting. They were found to be particularly useful, since they only have a theoretical understanding of problems in managerial accounting and do not have that much relevant work experience.

Abbreviations

ANOVA	Analysis of Variance
AT	Advanced Technologies
BIT	Bedrijfsinformatietechnologie
BK	Bedrijfskunde
BSC	Balanced Scorecard
BSK	Bestuurskunde
CE	Civil Engineering
CW	Communicatiewetenschappen
Df	Degrees of freedom
GZW	Gezondheidswetenschappen
MAC	Management Accounting and Controlling
M.B.A.	Master of Business Administration
PSY	Psychology
TBK	Technische Bedrijfskunde

List of tables and figures

Table of contents

1 Introduction

This master thesis is about common measures bias in the balanced scorecard (BSC), and to what extent this biases the performance evaluation of divisional managers. Common measure bias in the BSC arises when organizations, with multiple and slightly different BSCs for each division, need to compare the divisional managers for bonuses or career development. The BSC consists of *common measures*, which are the measures that fit the organizational strategy, and *unique measures,* which are the measures that are tailored to the divisional strategy. Common measure bias has most often been explained as decision makers' unwillingness to incorporate the unique information because this information requires greater cognitive effort process (Lipe & Salterio, 2000).

This introduction describes the research background and comes up with the problem definition. Thereafter, the problem definition will be redefined into a main research question that will be studied in this thesis.

1.1. Research background

The balanced scorecard (BSC) has been introduced as an integrated, balanced approach to performance measurement and improvement in which multiple organizational goals are measured and managed simultaneously to produce the desired results (Kaplan & Norton, 1992). The BSC is developed because managers like to have some insight in performance measures other than the traditional financial performance measures; these financial performance measures are called the leading indicators. Kaplan & Norton were not the first to advocate the importance of non-financial measures. In the 1950s, a project team of General Electrics recommended that divisional measures should be measured by one financial and seven non-financial measures (Kaplan, 2010; Malina & Selto, 2001).

The central idea behind a BSC is that if the organizations tracks the right set of leading indicators and gives them proper importance weightings, then profits will inevitably follow (Merchant & Van der Stede, 2007). A distinguishing feature of the BSC is the number and diversity of its indicators; BSCs contain mostly between sixteen and twenty-eight leading and lagging measures with performance measures along four dimensions (Banker, Chang, & Pizzini, 2011). The financial performance measures are grouped into one single category, the financial perspective. The non-financial performance measures (lagging indicators) are grouped into three categories: customer perspective, internal business process, and learning and growth (Cardinaels & Van Veen-Dirks, 2010). The BSC originally provide answers to four main questions (Kaplan & Norton, 1992):

- Financial perspective: How do we look to shareholders?
- Customer perspective: How do customers see us?
- Internal perspective: What must we excel at?
- Learning and growth perspective: Can we continue to improve and create value?

The combination of these four financial and non-financial performance measures is linked to the organization's vision and mission, and therefore linked to its strategy. Kaplan & Norton (2001) state: "*The best balanced scorecards reflect the strategy of the organization"*. Communicating this vision and strategy is important to successful implementation of the BSC (Kaplan & Norton, 1992, 1996, 2001). Kaplan & Wisner (2009) suggested to add a fifth performance measure to the BSC, to overcome certain biases in the BSC when a specific non-traditional strategic objective was present.

Because the BSC has a large number of performance measures, it presents a complex task to a manager asked to use the scorecard to evaluate division's performance (Lipe & Salterio, 2002). This complexity causes biases for decision managers in evaluating division managers, and therefore it is questionable whether the BSC is a victim of its own success.

The literature identifies the most common biases in the BSC are the negativity bias (Wong-on-Wing, Guo, Li, & Yang, 2007), the non-ambiguity bias (Liedtka, Church, & Ray, 2008), and the common measure bias (Lipe & Salterio, 2000, 2002). The negativity bias is referred to as "*when equal measures of good and bad are present, however, the psychological effects of the bad ones outweigh those of the good ones."*(Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Kaplan, Petersen, & Samuels, 2012). Liedtka, Church, & Ray, (2008) argue that ambiguity intolerance also can bias the BSC, this means that ambiguity-intolerant individuals are more likely to discount or ignore ambiguous information when the ambiguity relates to positive information (Liedtka, et al., 2008). Although these biases are an important field of study, this current study is focused upon the common measure bias.

Psychological research suggests that decision makers faced with both common and unique measures may place more weight on common measures than unique measures (Lipe & Salterio, 2000; Slovic & MacPhillamy, 1974). Lipe & Salterio (2000) has expanded this research towards the BSC, and studied whether common measures dominate BSC-based evaluations of subordinate units. They have conducted an experiment where M.B.A. students evaluate divisions of a clothing firm. The BSCs of these divisions are manipulated in the experiment to test whether this affects performance evaluation. The results from this study show that the experimental participants evaluate the divisions solely on the common measures. Performance on the unique measures do not affect the evaluation judgments (Lipe & Salterio, 2000).

1.2. Research problem

In reviewing the literature, many experimental studies have used M.B.A. students with a minimum of five years of work experience (Banker, Chang, & Pizzini, 2004; Banker, et al., 2011; Humphreys & Trotman, 2011; Lipe & Salterio, 2000). However, these M.B.A. students do not necessarily have prior work experience with the BSC, which causes a lack of relevant knowledge of the BSC. This is an important limitation of the Lipe & Salterio-experiment; their experimental participants are novices in the use of the BSC. They found that it could be difficult to identify appropriate participants for the experiments. They have suggested to use employees of a particular BSC firm, employees in a cross-section of BSC firms, or people who have been trained in the use of a BSC (Lipe & Salterio, 2000, p. 296). This limitation creates research opportunities for future research. Proponents of the BSC argue that training of participants is essential for successful implementation (Dilla & Steinbart, 2005a;

Niven, 2002). The level of knowledge and understanding of the BSC is likely to influence how decision makers use common and unique measures to evaluate divisional performance.

Sprinkle (2003) found that because organizations are relying more on both financial and non-financial measures, it seems vital to understand how and how well individuals understand these performance measures in evaluating divisional performance, and more general in making organizational desirable decisions. M.B.A. students are widely accepted as experimental participants for studies in managerial accounting. In general, M.B.A. students have a lot of full time work experience and are therefore expected to have a practical understanding of managerial accounting problems. In this research it is aimed to turn around that way of reasoning in using undergraduate students as experimental participants. They are expected to only have a theoretical understanding of the concept and design of managerial accounting.

Thus, using participants with a theoretical understanding of the BSC concept and design extends the research of Lipe & Salterio. Dilla & Steinbart (2005) suggested using undergraduate students with (almost) the same basic level of knowledge about the BSC. They have trained their participants through lectures and readings and by developing actual BSC's for two different organizations (Dilla & Steinbart, 2005a). In this thesis, I will extend the line of thinking by studying the impact of undergraduate students with a basic theoretical understanding of the BSC and no relevant work experience with the BSC. Undergraduate students are university or college students who are studying for their first bachelor degree. This thesis refers to undergraduate students as second year students who attended the course Management, Accounting and Controlling (MAC).

Also, this thesis complements the research of Slovic & MacPhillamy (1974) who have used undergraduate students to examine whether common cue dimensions have a greater influence on comparative judgments than when they are unique to a particular alternative (Slovic & MacPhillamy, 1974). In this study, student volunteers were asked to predict a student's Grade Point Average (GPA) based on numerical information in English skills, need for achievement, or quantitative skills. In comparing the students both students were judged on their English skills, one student was judged on need for achievement, and one student was judged on quantitative skills. It was found that when volunteers were asked to choose which student had the higher GPA and the degree of difference between the students, volunteers weighted information common to each pair of students more heavily than the unique measure (Slovic & MacPhillamy, 1974).

Based on this review above, I predict that common measure bias holds in an experiment with undergraduate students, and therefore the research question is:

Does common measure bias in the balanced scorecard hold in an experiment with undergraduate students?

1.3. Research outline

As a result from this problem definition, the goal of this research is twofold. First, to review the literature, there are already some leading experiments carried out with respect to the common measure bias. It is aimed to construct an overview of those experiments in order to provide a clear view of the existing literature on common measure bias. Second, to analyze data derived from the experiment with undergraduate students. This will be done carrying out a statistical analysis of the collected data from the experiment.

In the remaining of this research, the seven major experiments regarding common measure bias are reviewed. This literature review identifies five factors, which could attenuate common measure bias in the BSC. On the basis of these factors the hypotheses are formulated (Chapter 2). The methodological part describes and explains the experimental design of the study (Chapter 3). Based on the outcomes of the experiment, findings are presented. These findings are specified towards the four hypotheses (Chapter 4). Afterwards, the overall conclusion per hypothesis is given and whether the particular hypothesis is supported or not (Chapter 5). Finally, the findings are discussed; implications of the findings, and suggestions for future research are described. Also, a post hoc analysis on the multiple choice-questions is given (Chapter 6). In the appendices, the case materials are given and additional analyses on the findings in chapter 4 are conducted.

2 Literature review

This literature review provides insight in the balanced scorecard and focuses upon the description and exploration of the experiments carried out on the topic of common measure bias. Thus, the starting point of this literature review is the article of Lipe & Salterio, which is extensively elaborated in the upcoming section. Thereafter, the major successive experiments on attenuating the common measure bias in the balanced scorecard were searched by means of the snowball method. All of the articles discussed have cited Lipe & Salterio (2000) and therefore this literature review is limited to the studies that aim to replicate the experiment of Lipe & Salterio (2000).

> *Underuse of unique measures reduces the potential benefits of the BSC because the unique measures are important in capturing the unit's business strategy. (Lipe & Salterio, 2000)*

Lipe & Salterio were the first to emphasize the existence of common measure bias in the BSC. Common measure bias is the decision makers' unwillingness to incorporate the unique information because this information requires greater cognitive effort to process (Lipe & Salterio, 2000; Slovic & MacPhillamy, 1974). Slovic & MacPhillamy (1974) explored how undergraduate students use common and unique information when comparing two college students on their Grade Point Average (GPA). In a series of five experiments, the participants in their study compared two college students on one common measure and one unique measure. It was found that across all five experiments, the common measure slightly dominated the unique measure. Neither cautioning the students not to increase the weight of the common measures, nor feedback with the correct answer reduced the common measure bias in the experiment. Thus, it was found that cue dimension would have greater influence on comparative judgments on comparative judgments when they are common to a particular alternative than when they are unique to a particular alternative (Slovic & MacPhillamy, 1974).

Lipe & Salterio have applied these findings in their study, to test whether the BSC is due to common measure bias. Due to the fact that the BSC is ideally linked to each division's strategy and goals, it is not possible to create an overall BSC for the whole organization. However, some of the divisions' objectives are common with the objectives of other divisions. It is cognitively easier to compare the divisions on the basis of those common performance measures, which will result in common measure bias (Lipe & Salterio, 2000). Common measures often tend to be lagging and financial indicators of performance, whereas unique measures are often more leading and non-financial (Lipe & Salterio, 2000). Thus, managers may pay insufficient attention to leading and non-financial measures. This limited use or non-use of unique measures can have serious implications for business unit performance evaluation by managers (Lipe & Salterio, 2000).

In order to test whether common measure bias exists in the balanced scorecard, Lipe & Salterio have conducted an experiment among fifty-eight first year M.B.A. students. These students have on average a work experience of more than five years. They found that these M.B.A. students evaluated the performance of two business divisions solely on the measures common across the two divisions and that unique measures specific to each division had no effect on performance evaluation. The results of this study indicate that common measures across the two business divisions played a major role in performance evaluations, but unique measures did not have much influence on performance evaluation. The result of this analysis supports the finding of Slovic & MacPhillamy (1974) of a natural simplifying strategy. This simplifying strategy, known as *common measure bias*, results in managerial performance evaluations that reflect a greater weighting of measures common to two divisions and less weighting of measures unique to each division (Humphreys & Trotman, 2011).

With regard to the results from these two studies, it is questionable what reasons have caused managers to use more common measures than unique measures in performance evaluation. Successive research has tried to explore factors that may have caused common measure bias. Roberts, Albright, & Hibbets (2004) believed that the surplus of information could be the reason for not taking the unique measures into account in evaluating the performance of managers. They have found that if the balanced scorecard was disaggregated, common measure bias attenuates. Since the balanced scorecard typically contains four to seven performance measures in *each* of the four categories, disaggregation may be helpful to attenuate common measure bias. Disaggregated in this context means that participants in the experiment did not compare the divisional managers with each other, but instead assessed each division and its manager individually (Roberts, et al., 2004). Disaggregated judgment strategies are more advantageous the more complex the judgment required (Roberts, et al., 2004; Lyness and Cornelius, 1982). Furthermore, it was found that superiors appear to use the performance evaluations as part of their judgment for assigning bonuses, but they are inconsistent in applying the balanced scorecard as a standard for bonus allocation (Roberts, et al., 2004).

Dilla & Steinbart (2005) predict that knowledgeable decision makers will attend to both common and unique measures when bonus allocation (compensation) is linked to performance evaluation. They examine two measures of judgment quality: *consistency* between individual performance evaluation and compensation decisions and consensus among users' performance evaluation decisions. Both performance evaluation and compensation decisions are likely to involve comparisons across divisions, and therefore may be made in a similar manner (Dilla & Steinbart, 2005). Thus, they investigated the influence of training, experience and bonus allocation on performance evaluation. It was found that knowledgeable decision makers would use both common and unique measures when making bonus allocations, but still place greater weight on common measures than unique measures in performance evaluations.

Dilla & Steinbart (2005a) also believed that training of participants also attenuates common measure bias. They have found that providing supplementary information may improve judgment consistency and consensus by making it easier to compare individual BSC measures across divisions (Dilla & Steinbart, 2005a). The relative benefit of graphs versus tables depends on the nature of the decision task: graphs are more useful for tasks that require identifying and understanding relationships and for making comparisons, while tables are more useful for tasks that require extracting specific values and combining them into an overall judgment.

Libby, Salterio & Webb (2004) examines whether the common measure bias is effort-related or relates to data-quality concerns. They have provided their respondents a third-party assurance report to examine whether common measure bias relates to data-quality concerns. Assurance reports are valuable in decision-making because they enhance the reliability of data (Libby et al., 2004). To examine whether the common measure bias is effort-related, managers are required to provide a written report on their justification to their superior. It was found that participants who receive an assurance report believe all the performance measures are more relevant and reliable than those who do not receive such a report (Libby, Salterio, & Webb, 2004).

Banker, Chang & Pizzini (2004) have expanded the research design of Lipe & Salterio with strategically linked and non-linked measures. Thus, one division can outperform the other on the basis of as well common and unique measures, as well linked and non-linked measures, or a combination of them. They also account for the amount of strategy information the participants receive (no detailed information, or narrative and graphical information of the divisional strategy). It was found that participants who receive detailed strategy information perceive linked measures to be more useful than non-linked measures. With respect to the common and unique measures, if detailed information about the business strategy was provided to the participants the common measures still have a statistically significantly greater impact on performance evaluation than unique measures.

In a follow-up experiment, Banker, Chang & Pizzini (2011) found that participants who received strategy maps placed more weight on measures linked to strategy than participants who received only narrative strategy descriptions. Strategy maps are causal maps showing relations between BSC performance measures and overriding strategic objectives (Banker, et al, 2011). They can aid managerial decisions if they enable managers to assess a measure’s relative importance to the achievement of strategic goals and thus provide indications for managers to weight and aggregate BSC measures in formulating an overall decision.

In addition to the research of Banker, Chang & Pizzini (2004, 2011), Humphreys & Trotman (2011) also investigated the role of providing strategy information to participants. They found that when participants are given strategy information (including a strategy map), and if all measures are strategically linked, common measure bias is attenuated, but if only half the measures are linked the bias remains (Humphreys & Trotman, 2011).

As can be derived from the review above, five factors that may attenuate common measure bias in the balanced scorecard. It was found that common measure bias could be attenuated if participants:

1. Have to disaggregate the balanced scorecard (e.g. Roberts, Albright, & Hibbets, 2004),
2. Have to link performance evaluation and compensation (e.g. Dilla & Steinbart, 2005a)
3. Are trained in the use of the balanced scorecard (e.g. Dilla & Steinbart, 2005)
4. Have to provide a written assurance report about the judgment (e.g. Libby, et al., 2004)
5. Receive detailed strategy information (Banker, et al., 2004, 2011; Humphreys, et al., 2011)

In this thesis, the first two factors will be further investigated. Before continuing investigating these two factors, an overview of all the experiments discussed above is given in table 2.1.

Study	*Dependent variable (Y)*	*Between-subjects variables*	*Within-subjects factor*	*Participants*	*Work experience of participants*	*Common measure bias attenuates by:*
Lipe & Salterio (2000)	Performance evaluation *[0 = Reassign – 100 = Excellent]*	1. Relative performance on common measures **(2)** 2. Relative performance on unique measures **(2)**	Division (RadWear and WorkWear	58 first year M.B.A. students of which 63% is male	More than 5 years of work experience	**No attenuation, only recognition of common measure bias in the balanced scorecard.**
Roberts, Albright & Hibbets (2004)	Performance evaluation *[0 = Reassign – 100 = Excellent]*	1. Relative performance on common measures **(2)** 2. Relative performance on unique measures **(2)**	Division (RadWear and WorkWear)	79 M.B.A. students, of which 25 were Executive M.B.A. students, and 54 were regular M.B.A. students	5,1 years of work experience	Disaggregation of the balanced scorecard
	Bonus allocation of $100.000					
Banker, Chang & Pizzini (2004)	Performance evaluation *[0 = Reassign – 12 = Excellent]*	1. Outperformance on common (unique) and/or linked (non-linked) measures **(16)** 2. Strategy information (yes/no) **(2)**	Division (The Women's Store and The Family Store)	480 M.B.A. students	6,4 years of full-time work experience	Strategy maps

Study	*Dependent variable (Y)*	*Between-subjects variables*	*Within-subjects factor*	*Participants*	*Work experience of participants*	*Common measure bias attenuates by:*
Libby, Salterio & Webb (2004)	Performance evaluation *[0 = Reassign – 100 = Excellent]*	1. Provision (or not) of a third party assurance report **(2)** 2. Requirement (or not) to justify their evaluation of each manager's performance **(2)** 3. Mixed crossing: RadWear scores better on common measures while WorkWear scores better on unique measures. **(2)**	Division (RadWear and WorkWear)	227 M.B.A. students from four public universities	5,8 years of full-time work experience	Third party assurance report and justification of performance evaluation.
Dilla & Steinbart (2005)	Performance evaluation *[0 = Reassign – 100 = Excellent]* Bonus allocation of $20.000	1. Information display **(3)** - Divisional BSC only - Supplemental tables - Supplemental graphs 2. Relative performance on common measures **(2)** 3. Relative performance on unique measures **(2)**	Division (RadWear and WorkWear)	132 undergraduate students	45% of the participants had more than one year of full-time work experience	Training of participants

Study	*Dependent variable (Y)*	*Between-subjects variables*	*Within-subjects factor*	*Participants*	*Work experience of participants*	*Common measure bias attenuates by:*
Dilla & Steinbart (2005a)	Performance evaluation *[0 = Reassign – 12 = Excellent]*	1. Relative performance on common measures **(2)** 2. Relative performance on unique measures **(2)**	Division (RadWear and WorkWear)	43 undergraduate students	3,1 years of full-time	Linking performance evaluation to compensation decisions
	Bonus allocation of $20.000					
Humphreys & Trotman (2011) [Experiment 1]	Performance evaluation *[0 = Reassign – 100 = Excellent]*	1. Extent of strategic linkage **(2)** - Half linked / half non-linked to divisional strategy - Fully linked to divisional strategy 2. Provision of strategy information **(2)** - No strategy information - Detailed strategy information 3. Divisional out performance pattern **(2)**	Division (General Jeans and Captain Kids)	92 Executive M.B.A. students	11,2 years of full-time work experience	Strategy maps

Study	*Dependent variable (Y)*	*Between-subjects variables*	*Within-subjects factor*	*Participants*	*Work experience of participants*	*Common measure bias attenuates by:*
Humphreys & Trotman (2011) [Experiment 2]	Performance evaluation *[0 = Reassign – 100 = Excellent]*	1. Focus on common theme **(2)** - Asset productivity-focused common theme - Profit-focused common theme 2. Provision of strategy information **(2)** - No strategy information - Detailed strategy information 3. Divisional out performance pattern **(2)**	Division (General Jeans and Captain Kids)	103 M.B.A. students	10,9 years of full-time work experience	Strategy maps
Banker, Chang & Pizzini (2011)	Performance evaluation *[0 = Reassign – 12 = Excellent]*	1. Relative performance on common (unique) and/or linked (non-linked) measures **(16)** 2. Strategy information **(3)** - No strategy information - Strategy map - Narrative information	Division (The Women's Store and The Family Store)	180 M.B.A. students	7 years of full-time work experience	Strategy maps

Table 2.1. Overview of the experiments carried out on the topic of common measure bias

If the balanced scorecard is disaggregated, will this influence the performance evaluation of undergraduate students? If compensation decisions were linked to performance evaluation, would the compensation decisions be affected by performance evaluation?

In conclusion, it was found that familiarity with the BSC would attenuate common measure bias. Based upon the findings of the experiments discussed above, in this study it is aimed to investigate whether common measure bias as found in the study of Lipe & Salterio (2000) is also present in a study with undergraduate students. Therefore, the *main hypothesis* of this study is formulated as follows:

H1: Common measure bias is also found in an experiment with undergraduate students.

As stated before, this study will be continuing investigating whether disaggregation strategies will attenuate common measure bias in the BSC and whether compensation decisions are affected by performance evaluation using a BSC.

Linking disaggregation strategies to performance evaluation

Another point of interest is the disaggregation of the balanced scorecard. Libby & Libby (1989) found that judgments are less than perfectly consistent and important cues are often ignored or misweighted. They found that disaggregating decisions increases consensus and inter-judge agreements (Roberts, et al, 2004; Libby & Libby, 1989). Roberts, Albright & Hibbets (2004) extended these findings towards the balanced scorecard. They found that participants made use of twice the number of BSC items as Lipe & Salterio's participants when disaggregating the balanced scorecard (Roberts, et al., 2004).

In this study it is aimed to test whether disaggregation of the balanced scorecard influences undergraduate students in the performance evaluation of divisional managers. It is therefore hypothesized that:

H2: Disaggregation of the performance evaluation will attenuate common measure bias in the balanced scorecard in an experiment with undergraduate students.

Linking performance evaluation to compensation decisions

Kaplan & Norton (2001b) suggested that the BSC could also be used to make compensation decisions. Compensation refers to the exchange of service and rewards between employees and organizations (DoHerty & Nord, 1993). Traditionally, organizations have measured and rewarded managerial performance only using financial performance measures (Banker, Potter, & Srinisivan, 2000). Banker, Potter, & Srinisivan (2000) found that the inclusion of non-financial performance measures in compensation contracts would enhance the overall performance of managers. In addition, Ullrich & Tuttle (2004) found that using nonfinancial measures in compensation decisions caused managers to spend more attention to the nonfinancial areas. Kaplan & Norton (1996, 2001b) suggested that organizations should link the compensation decisions to the BSC to enhance the usefulness of the BSC. However, while the BSC and the compensation system are closely tied, they are not the same.

In this study it is aimed to test whether compensation decisions of undergraduate students are affected by performance evaluations. Roberts, Albright & Hibbets (2004) found that decision makers appear to use the balanced scorecard as part of their judgment models for compensation decisions, but they are inconsistent in their application of the information from the BSC. In this study, two different types of compensation decisions were used, bonus allocation and promotion. The performance evaluation scores are an input for compensation decisions. These compensation decisions were made separately from the performance evaluations. It is therefore hypothesized that:

H3: Performance evaluations based on a balanced scorecard will affect subsequent bonus allocations in an experiment with undergraduate students.

H4: Performance evaluations based on a balanced scorecard will affect subsequent promotion decisions in an experiment with undergraduate students.

3 Methodology

This chapter explains the methodological part of this study. In order to answer the research question as formulated in chapter 1 (Does common measure bias in the balanced scorecard hold in an experiment with undergraduate students?), the hypotheses as formulated in the previous chapter have been tested by means of an experiment. As explained in the previous chapter, the experiment of Lipe & Salterio (2000) is leading in the literature about common measure bias. To maximize the comparability with their study, the experimental design of this thesis is based upon that of Lipe & Salterio (2000).

3.1. Participants

The participants in this experiment are second-year students who have followed the course Management Accounting and Controlling (MAC) during the fourth quartile of the academic year 2010-2011. Data is collected at the end of the course MAC on June 9th, 2011. The sample exists of two hundred and nine (209) students who have followed the course MAC. The participation in the experiment is voluntary. During the lectures of MAC, the students were trained in the use of the balanced scorecard. The following time path indicates the training effect at MAC:

May 24, 2011:	Lecture on advantages of financial control
May 30, 2011:	Lecture on problem of myopic behavior of financial control
May 31, 2011:	Lecture on the BSC / Strategy maps
June 6, 2011:	Guest lecture of Imtech BSC consultant
June 9, 2011:	Experiment (small part of the exam).

There were 209 participants in the experiment. Only the responses of two hundred and seven (207) students are used in this study because one student failed to complete the performance evaluation for both managers, and another student did not complete the control part of the experiment. From the 207 experimental participants, 71 were female (34.3%) and 136 were male (65.7%). The students do not have much full-time work experience (0.17 years on average, with a maximum of five years). Table 3.1 shows the descriptive statistics of the experimental participants.

A t-test indicates that the work experience of the participants in this study is significantly less (t=-42.472, p<0.01) than the work experience of the participants in the experiment of Lipe & Salterio (2000), in which the students have on average a work experience of five years. The average age of the experimental participants in the study of Lipe & Salterio (2000) is not given. In the study of Lipe & Salterio (2000), 63% of the experimental participants were male. This ratio 'male-female' is almost equal to that of this current study. So, the students have less work experience and are expected to only have a theoretical understanding of the concept and design of the BSC. So, the sample of experimental participants is appropriate to study.

Table 3.1. Descriptive statistics of the undergraduate students participating in the experiments (n = 207)

	Mean	
Age (years)	21.2	
Full-time work experience (years)	0.17	
Gender – number (%) female	71 (34,3%)	
Number of relevant courses BSC	1.18	
Study	*Number of participants*	*%*
Bedrijfskunde (BK)	69	33.33%
Technische bedrijfskunde (TBK)	45	21.74%
Bestuurskunde (BSK)	36	17.40%
Gezondheidswetenschappen (GZW)	32	15.46%
Bedrijfsinformatietechnologie (BIT)	20	9.66%
Psychology (PSY)	2	0.97%
Advanced Technologies (AT)	1	0.48%
Civil Engineering (CE)	1	0.48%
Communicatiewetenschappen (CW)	1	0.48%
Total	**207**	**100.00%**
Area of full-time work experience	*Number of participants*	*%*
Accounting, auditing or taxation	1	0.48%
Marketing or sales	9	4.35%
Retail industry	5	2.43%
Other	6	2.90%
No full-time work experience	187	90.34%
Total	**207**	**100.00%**
Experience in retail-clothing store	*Number of participants*	*%*
Less than a month or during one school holiday	1	0.48%
Part time during a school year	1	0.48%
Part time during multiple school years	4	1.94%
No work experience in a retail-clothing store	201	97.10%
Total	**207**	**100.00%**

3.2. Experimental design

Lipe & Salterio (2000) developed their experimental case materials on the basis of the Kenyon Stores case of Kaplan & Norton (1996). The case introduces the mission statement: *“We will be an outstanding apparel supplier in each of the specialty niches served by WCS.”* (Lipe & Salterio, 2000, p. 288), the divisional strategies of RadWear and WorkWear, and presents the balanced scorecards of its two major divisions, RadWear and WorkWear. The divisional strategy of RadWear was given as (Lipe & Salterio, 2000, p. 289):

> *RadWear’s management determined that its growth must take place through an aggressive strategy of opening new stores. RadWear also determined that it must increase the number of brands offered to keep the attention and capture the clothing dollars of its teenage customers. RadWear concluded that its competition radius is fairly small due to the low mobility of young teens.*

The divisional strategy of WorkWear was given as (Lipe & Salterio, 2000, p. 290):

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

On the basis of this information the participants were asked to evaluate the performance of the managers on a 101-point scale [0 = Reassign – 100 = Excellent].

```
   0                            50                           100
   | - - - | - - - | - - - | - - - | - - - | - - - | - - - | - - - | - - - | - - - |
Reassign  Very       Poor   Average      Good    Very     Excellent
          Poor                                   Good
```

The experiments subsequent to that of Lipe & Salterio (2000) used a similar scale, also with seven descriptive labels. The performance evaluation of the managers is in all experiments the dependent variable. However, Roberts, Albright & Hibbets (2004) and Dilla & Steinbart (2005) have added a second dependent variable to their experimental design. The participants were asked to allocate a bonus of $20.000 or $100.000, respectively.

Despite the fact that all of the experiments have used outperformance on the common and unique measures as between-subject factors, there is a large variety in additional between-subject factors. These additional between-subject factors are incorporated since it is for the purpose of that particular study. For example, Libby, Salterio & Webb (2004) have incorporated “justification” and “providing an assurance report” in their between-subjects design to test whether this has an effect on common measure bias or not.

With respect to the within-subjects factor, all of the experiments have used the same factor, to maximize the comparability of the studies. The participants have to assign a score to the divisional managers of two divisions of an organization in the retail industry, one division in a mature business stage and the other in a high-growth business stage. Most of the studies have adopted the case of Lipe & Salterio (2000), with the divisions called RadWear and WorkWear.

Experimental participants were asked to act as a senior executive (superior) assigning scores to the two divisional managers on the basis of the provided BSC. Most of the studies have used M.B.A. students as experimental participants, mostly having more than five years of full-time work experience. In contrast, Dilla & Steinbart (2005, 2005a) have used undergraduate students as experimental experiments, with almost no full-time work experience. They argue that training of participants could ensure that they have a thorough understanding of the concept of the BSC.

This experiment differs in many ways from the previous experiments on common measure bias. Compared to the study of Lipe & Salterio (2000) an important difference is the use of undergraduate students instead of M.B.A. students with a work experience of more than five years. While the use of M.B.A. students as experimental participants is widely accepted, the use of undergraduate students as experimental participants is particularly new in managerial accounting. Despite the fact that the students in this experiment do not have much full-time work experience, I believe that they are appropriate to investigate. The students were trained in the use of the balanced scorecard, and are therefore expected to have a theoretical understanding of the concept and design of the BSC. It is appropriate to study undergraduate students because this paves the way for future research on managerial accounting with undergraduate students.

Second, in this experiment there are sixteen different BSCs developed, whereas Lipe & Salterio (2000) only used one standardized BSC. Another differentiation from the original experiment is the use of a 13-point scale instead of a 101-point scale. Fourth, some of the students have to evaluate the performance of RadWear first, while others have to evaluate the performance of WorkWear first. The experiment of Lipe & Salterio knows a standard order of the BSCs. Also the disaggregation of the BSC and the bonus allocation / promotion is not present in the experiment of Lipe & Salterio. The complete experimental design of this current study can be presented as follows:

*Performance evaluation (16) * Bonus allocation (2) * Aggregation (2) * Order (2)*

This means that there are one hundred and twenty-eight (128) different treatments in this experimental design. In a balanced design, this amount of experimental participants is multiplied, which makes it possible to compare the outcomes of multiple students in exactly the same experimental condition, with each other.

Since it is aimed to replicate the experiment of Lipe & Salterio (2000), the analyses in the next chapter are based upon the responses of 56 students.[1] These are the students who are in the experimental conditions equal to that of Lipe & Salterio (2000). They have received a BSC in which

[1] Appendix VI presents additional results on the responses of all of the 207 experimental participants.

RadWear or WorkWear is favored on common or unique measures. There are also students who received a BSC in which no division is favored. The analysis is based upon the responses of students who received a BSC in which RadWear or WorkWear is favored on common or unique measures.

For hypothesis 2, stating that disaggregation of the performance evaluation will attenuate common measure bias in the balanced scorecard in an experiment with undergraduate students; the distribution of students across the different experimental conditions is given in table 3.2.

	Mechanically aggregated score	Holistic score	Total:
Total sample:	**101**	**106**	**207**
Favor RadWear or WorkWear	*27*	*29*	*56*
No division favored:	74	77	151

Table 3.2. Distribution of students across experimental conditions for disaggregation strategies

For hypotheses 3 and 4, stating that bonus allocations / promotion decisions of undergraduate students are affected by performance evaluation using a balanced scorecard; the distribution of students across the different experimental conditions is given in table 3.3.

	Bonus allocation	Promotion	Total:
Total sample:	**117**	**90**	**207**
Favor RadWear or WorkWear:	*31*	*25*	*56*
No division favored:	86	59	151

Table 3.3. Distribution of students across experimental conditions for compensation decisions.

3.2.1. Between-subjects design

27 students rated the managers of both divisions on each of the sixteen individual items of the BSC. After completing this performance evaluation of both managers, the students were asked to multiply those individual judgments by pre-determined weights and to sum those weighted scores to calculate the total weighted score of the managers, the *mechanically aggregated score*. When the students disagree with the total weighted score of a manager, they have the possibility to adjust this judgment if they were not satisfied with the outcome of their mechanically aggregated score for any reason (Roberts, et al., 2004). The other 29 students only have to assign a holistic score to both managers. It is expected that disaggregation of the BSC will attenuate common measure bias in the BSC (Roberts, et al., 2004).

The next condition in the experiment is the order of the BSCs. Some students first have to judge Chris Peeters (RadWear), while others first have to judge Bob Graham (WorkWear). It is expected that *order* does not make a difference in the judgment of the managers (Humphreys & Trotman, 2011).

After completing the first part of the experiment, the students were asked to seal the envelope with the performance evaluation and hand it in at the examiner. So, they have no access to the case materials and the balanced scorecards of both divisions when filling in the second part of the experiment. This second part contains among others the bonus allocation-part.

3.2.2. Within-subjects factor

The within-subjects factor is the division to be evaluated. The participants needed to evaluate the performance of Chris Peeters (RadWear) and Bob Graham (WorkWear). This within-subjects factor is used to test whether there are differences in performance evaluations when (1) the balanced scorecard is disaggregated or not, or (2) the order of the balanced scorecard is reversed. As it is hypothesized that disaggregation of the balanced scorecard will influence the performance evaluation of undergraduate students, it is expected that the first condition will cause differences in the performance scores of the two divisional managers. However, it is expected that the order of the balanced scorecards of the two divisions do not cause differences in the performance scores.

3.2.3. Dependent variable

The first dependent variable is the difference in performance evaluation of the two divisional managers. The scores assigned to the divisional managers by the participants measure the performance evaluation of the two managers. The students ranked the divisional managers on a 13-point scale, adopted from Banker, Chang, and Pizzini (2011). The advantage of using a 13-point scale (instead of a 101-point scale) is that it makes it easier for the respondent to give a more accurate score to the manager, because both the scales of Lipe & Salterio (2000) and that of Banker, Chang & Pizzini (2011) have only seven descriptive labels. The descriptive labels of the scale are:

12 = Excellent: far beyond expectations, manager excels
10 = Very good: considerably above expectations
8 = Good: somewhat above expectations
6 = Average: meets expectations
4 = Poor: somewhat below expectations, needs some improvement
2 = Very poor: considerably below expectations
0 = Reassign: sufficient improvement unlikely

The second dependent variable in this case is the bonus allocation of $100,000. In the second part, 31 students have to allocate a bonus of $100,000 between Chris Peeters (RadWear) and Bob Graham (WorkWear). This dependent variable is used to test whether bonus allocation is affected by performance evaluation based on a BSC.

25 students made a separate overall assessment of the managers performance, measured on a 13-point scale as defined above [0 = Reassign – 12 = Excellent]. This dependent variable is used to recommend a promotion to one of the two divisional managers, and to test whether promotion decisions are affected by performance evaluation based on a BSC.

3.2.4. Manipulation of the balanced scorecard

The BSCs of both RadWear and WorkWear contain sixteen performance measures, four in each category (financial, customer, internal, learning & growth). Each category consists of two common measures and two unique measures. An overview of the performance measures used in the BSC is given in table 3.2.

An example of RadWear and WorkWear's BSC is given in appendix III. In these tables, showing the BSCs of both divisions, the columns "Actual", "Target" and "%better than target" appear. The percentages / numbers in the "Actual" column represents the actual performance on that measure. The percentages / numbers in the "Target" column represents the target on that performance measure. The last column "%better than target" reflects the difference between actual and targets.

In each category (financial, customer, internal, learning & growth) RadWear could be superior on common measures (COM-Rad) or WorkWear could be superior on the common measures (COM-Work). The other way around, RadWear could be superior on the unique measures (UNIQ-Rad) or vice versa. There are also BSCs in which none of the two outperform the other on common or unique measures. These sixteen different scorecard versions reflect each of the sixteen combinations possible. It is expected that this manipulation will cause differences in the performance scores of the divisional managers.

Table 3.4. Performance measures used in the balanced scorecard (source: Lipe & Salterio, 2000)

Type	*Measure*
Financial measures	
Common	Return on sales
Common	Sales growth
Unique (RadWear)	New store sales
Unique (RadWear)	Market share relative to retail space
Unique (WorkWear)	Revenues per sales visit
Unique (WorkWear)	Online sales
Customer measures	
Common	Repeat sales
Common	Customer satisfaction rating
Unique (RadWear)	Mystery shopper program rating
Unique (RadWear)	Returns by customers as % of sales
Unique (WorkWear)	Captured customers
Unique (WorkWear)	Referrals
Internal business measures	
Common	Returns to suppliers
Common	Average markdowns
Unique (RadWear)	Average major brand names/store
Unique (RadWear)	Sales from new market leaders
Unique (WorkWear)	Orders filled in one week
Unique (WorkWear)	Web orders filled without errors
Learning & growth measures	
Common	Hours of employee training / employee
Common	Employee suggestions / employee
Unique (RadWear)	Average tenure of sales personnel
Unique (RadWear)	Stores computerizing
Unique (WorkWear)	% Staff with M.B.A. degrees
Unique (WorkWear)	Database certification of clerks

3.3. Procedures

The undergraduate students were asked to read the case of Chadwick Ltd., a clothing retailer that recently implemented the balanced scorecard within the organization.[2] In the case, two divisions of Chadwick Ltd. were discussed, RadWear and WorkWear, with a different strategy and target market. The case was adopted from the study of Lipe & Salterio, and the participants were asked to fulfill the role of a senior executive of Chadwick Ltd. The case informed the students about the mission statement and strategic goals of both divisions and introduced them to the divisional managers.

Thereafter, the balanced scorecards of the two divisions were provided to the students and their experimental tasks were explained. After finishing the performance evaluation and handing it in, the students answered some questions with demographic information, responded to ten statements with manipulation checks (appendix IV), and answered twelve questions to check whether they understood the concept of the balanced scorecard. The debriefing questionnaire with demographic information asks for age, gender, study, work experience, experience with the balanced scorecard and so on. This information indicates whether the participants fit within the scope of undergraduate students.

3.4. Data analysis

In order to maximize the comparability of the results, much of the structure of the experiment of Lipe & Salterio (2000) is replicated. To test whether common measure bias holds in an experiment with undergraduate students (hypothesis 1), at first a repeated measures ANOVA (2 x 2 x 2) on divisional performance is conducted with division as the repeated measure and common and unique measures as the between-subject factors. Repeated measures accounts for the fact that two measurements were taken (one for each division) from each participant. The interaction effects between the between-subjects factors and the within-subjects factor are tested. Thus it is tested whether there is interaction between division and common and/or unique measures. Also, a regression analysis is conducted to further investigate the relative influence of common and unique measures. In this analysis, the dependent variable is the difference in the performance evaluations of the two divisional managers of RadWear and WorkWear. In order to compare the results of the Lipe & Salterio-experiment with the results of this experiment, the eta-squared (η^2) of the repeated measures ANOVA is calculated.

To test whether disaggregation influences performance evaluation of undergraduates students (hypothesis 2), repeated measures ANOVA (2 x 2 x 2) for both disaggregated and overall judgments are conducted. The results of these two analyses are compared to reject or accept the hypothesis.

To test whether bonus allocations are affected by performance evaluation using a BSC (hypothesis 3), a regression analysis on differences in bonus allocations is conducted. Also the fourth hypothesis, stating that promotion decisions are affected by performance evaluation using a BSC, is tested by a regression analysis.

[2] Case materials are available from Dr. T. de Schryver.

3.5. Reliability

Experimentation involves the active and purposeful manipulation and measurement of variables, thereby enabling the researcher to create a research setting and generate data (Sprinkle, 2003). The unique strength of experimentation is in describing the consequences attributable to deliberately varying a treatment (Shadish, Cook, & Campbell, 2002).

With respect to reliability of the experimental design, though the participants are chosen carefully the experimental design have some limitations. On forehand, 223 participants have enrolled for the experiment, which is the amount of participants used to set up the experiment. Because the participation in the experiment was voluntary, not all of the enrolled participants were present. This causes differences in experimental groups for bonus allocation (90 students versus 117 students) and aggregation (101 students versus 106 students). However when looking at complete experimental design "bonus allocation * aggregation * order * performance evaluation", every of the 128 experimental conditions is completed by at least one student.

3.6. Manipulation checks

While the first part of the experiment consists of the BSCs and performance evaluation of divisional managers, the second part consists of various manipulation checks. Gravetter & Forzano (2011) found that manipulation checks are particularly useful in four situations: participant manipulations, subtle manipulations, simulations, or placebo control conditions. In this experiment a manipulation check with ten statements on the understanding of the case materials is conducted and it is tested whether order effects influences performance evaluation.

Understanding of the case materials

First, it is checked whether the undergraduate students have understood the case materials. This is measured on a scale from -5.0 to 5.0, with -5.0 indicating the participants strongly disagree with the statement and 5.0 indicating the participants strongly agree with the statement. This does not mean that all of the statements need a positive outcome. In the second column of table 3.3 the expected outcome of each specific statement is given. Except for the second and fourth statement, all the means are expected to be positive. Table 3.3 shows that all of the outcomes are answered as expected (on average). With respect to the experiment, the students found the case easy to understand (mean = 2.30), easy to do (mean = 1.23), and found the case realistic (mean = 2.43). All of the means were significantly different from zero ($p < 0.01$).

Manipulation check	*Expected outcome*	*Mean (SD)*
The two divisions, RadWear and WorkWear, use different performance measures.	Positive (+)	1.95 (2.027)***
The two divisions, RadWear and WorkWear, sell to the same markets	*Negative (-)*	*-3.89 (1.534)****
It was appropriate for RadWear and WorkWear to employ different performance measures	Positive (+)	2.46 (1.910)***
The strategy of RadWear is to generate greater sales through its existing infrastructure rather than to invest in new stores.	*Negative (-)*	*-3.10 (2.598)****
The strategy of WorkWear is to target the companies more than the employees.	Positive (+)	1.90 (2.316)***
To grow sales, RadWear must successfully introduce new lines of clothing to its existing customers	Positive (+)	2.30 (2.338)***
WorkWear relies on new distribution channels to retain existing customers	Positive (+)	1.43 (2.805)***
The case material was easy to understand.	Positive (+)	2.30 (1.913)***
The case was easy to do	Positive (+)	1.23 (2.209)***
The case material was realistic	Positive (+)	2.43 (1.713)***

Table 3.5. The mean scores of the respondents on whether they understood the case materials
**P<0.10; ** P< 0.05; ***P<0.01 (two-tailed)*

So, the experimental case materials are appropriate to use in this experiment because this manipulation check indicates that the undergraduate students understand the differences between the two divisions. Also, the undergraduate students do not perceive the case materials as difficult.

Order effects

Also, it was accounted for whether differences in the order of the presentation of the balanced scorecard, caused differences in the performance evaluation scores. Table 3.4 shows that neither the differences of "RadWear first" (t=-1.155, p=0.877) and "WorkWear first" (t=0.583, p=0.561) are statistically significant. This indicates that the mean impact of order effects do not influence performance evaluation scores.

Order	*Division*	*Mean*	*N*	*t-value*	*p*
RadWear first	RadWear	8.23	116	- 1.155	0.877
	WorkWear	8.25	116		
WorkWear first	RadWear	8.35	91	0.583	0.561
	WorkWear	8.28	91		

Table 3.6. Average performance evaluation scores corrected for order effects.

In conclusion to the methodological part of this study, it could be stated that the experimental participants are appropriate to study because they have significantly less work experience than the participants in the study of Lipe & Salterio (2000). This means that they only have a theoretical understanding of the concept and design of the BSC, which strengthens the findings of this study. The experimental design, as described in paragraph 3.2, is found to be appropriate because the manipulation checks do not notice anything unusual. The participants indicate that they understand the case materials, and order effects do not influence performance evaluation scores.

4. Analysis

In this chapter, the findings from the experiment were reviewed. One main hypothesis and three sub hypotheses were used to test whether undergraduate students could be used for research on attenuating common measure bias. The hypotheses will be explored on the basis of statistical analyses on the data from the experiment.

H1: Common measure bias is also found in an experiment with undergraduate students

Starting with the main hypothesis, two different situations were analyzed to test whether common measure bias is also found in an experiment with undergraduate students. In the first situation, the experiment of Lipe & Salterio is replicated. Second, the first analysis is adjusted for students who have adapted their mechanically aggregated score.

Situation 1: Replication of the Lipe & Salterio-experiment

First, a repeated measure ANOVA is used to test whether undergraduate students show the same degree of common measure bias as M.B.A. students. Since the between-subjects design of Lipe Salterio (2000) is a 2 x 2 design in which RadWear is favored *or* WorkWear is favored, and it is aimed to replicate their analysis, only 56 students are incorporated in the analyses. These are the students who are in the experimental condition equal to that of Lipe & Salterio (2000).

Table 4.1. Results of a 2 x 2 x 2 repeated measures ANOVA of evaluations of the performance of RadWear and WorkWear division managers (n = 56) [a]

Between-subjects	*df*	*Sum of squares*	*Mean Square*	*F-value*	*p*
Common	1	0.541	0.541	0.351	0.556
Unique	1	0.095	0.095	0.062	0.805
Common * Unique	1	6.077	6.077	3.940	0.052*
Error	52	80.194	1.542		

Within-subjects	*df*	*Sum of squares*	*Mean Square*	*F-value*	*p*
Division	1	0.149	0.149	0.306	0.583
Division * Common	1	10.926	10.926	22.420	0.000***
Division * Unique	1	6.154	6.154	12.628	0.001***
Division * Common * Unique	1	0.001	0.001	0.002	0.960
Error	52	80.194	1.542		

**P<0.10; ** P< 0.05; ***P<0.01*

[a] Evaluations made on a 13-point scale, with 0 labeled “Reassign” and 12 labeled “Excellent”

The results of this repeated measures ANOVA are given in table 4.1. It is found that both the *Division x Common* interaction effect (F=22.420, $p < 0.01$) and the *Division x Unique* interaction effect (F=12.628, $p < 0.01$) are significant. This means that in this experiment, both the common measures and the unique measures affect the managers' evaluations. This is in contrast with the study of Lipe & Salterio, who found only a significant interaction effect between division and common measures, which they defined as *common measure bias*.

Thus, in contrast to the findings of Lipe & Salterio (2000) undergraduate students use both common and unique measures in performance evaluation. So, no common measure bias is found. Therefore these results suggest that undergraduate students do not show the same degree of common measure bias compared to M.B.A. students.

Moreover, table 4.2 shows that when common measures favor RadWear, the manager of RadWear is evaluated 0.62 points higher on average than the manager of WorkWear. When the common measures favor WorkWear, that divisional manager is evaluated 0.55 points higher on average. In the same situation for the unique measures, the RadWear (WorkWear) manager is evaluated 0.67 (0.70) points higher on average. The mean difference in performance evaluation in both situations is close to zero (RadWear = 0.07; WorkWear = 0.03), and these differences are found to be non-significant.

Table 4.2. Evaluations of the Performance of RadWear and WorkWear Division Managers[b]

	RadWear	*WorkWear*	*Difference* *RadWear - WorkWear*
Common Measures			
Favor RadWear	8.49 (1.24)	7.87 (1.39)	0.62
Favor WorkWear	8.07 (0.87)	8.62 (1.22)	- 0.55
Unique measures			
Favor RadWear	8.50 (0.88)	7,83 (1,13)	0.67
Favor WorkWear	8.06 (0.90)	8,76 (1,14)	- 0.70

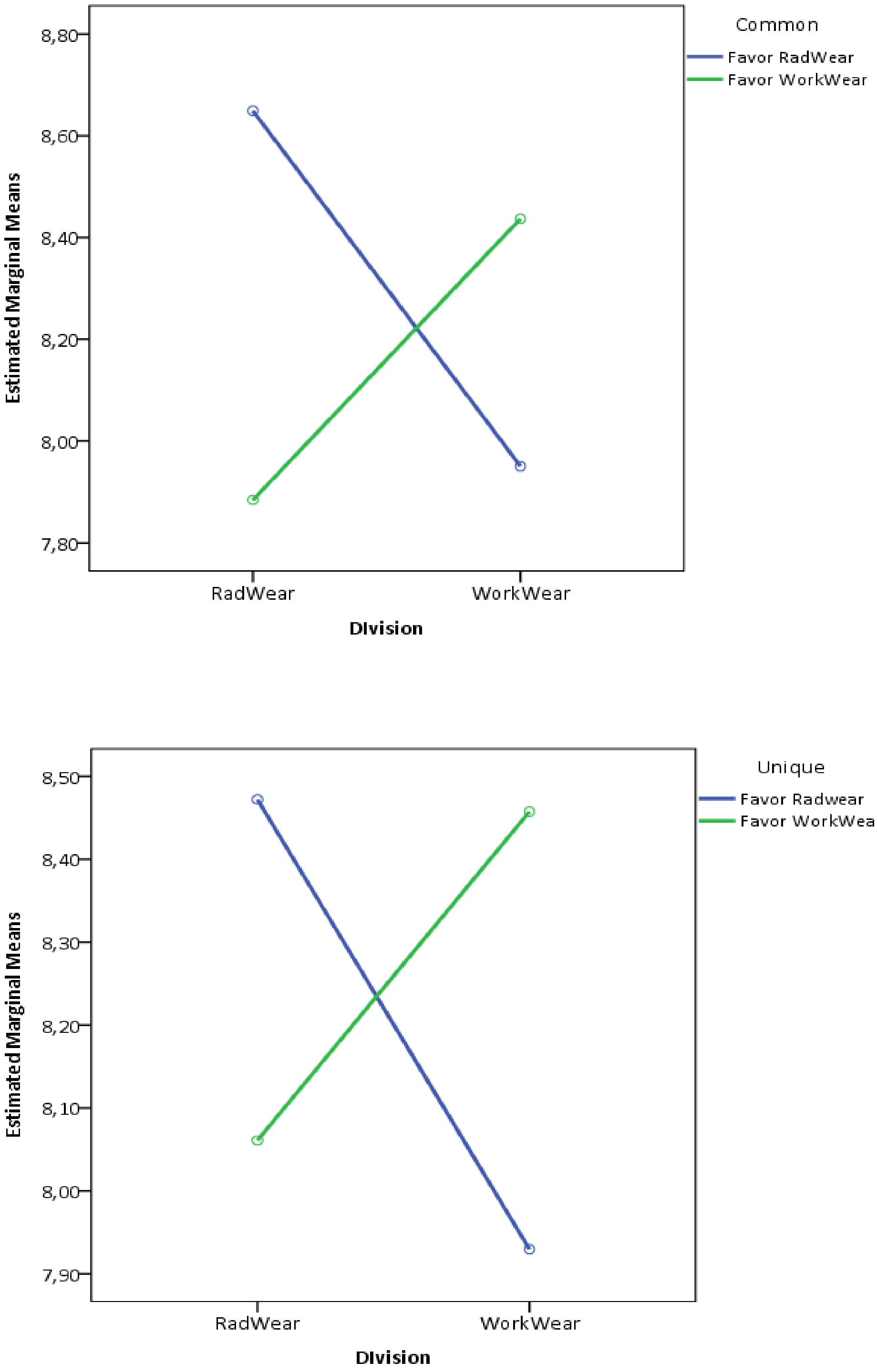
**P<0.10; ** P< 0.05; ***P<0.01*

[b] Table values are means (standard deviations). *Common* measures appear on both divisions' balanced scorecards. *Unique* measures appear on only one division's balanced scorecard. *Favor RadWear* indicates the measures were higher for the RadWear division than the WorkWear division. *Favor WorkWear* indicates the measures were higher for the WorkWear division than the RadWear division.

To compare the results of Lipe & Salterio (2000) with the results from table 4.1, the eta-squared (η^2) is calculated. The eta-squared is calculated using the following equation: $\eta^2 = SS_{between} / SS_{total}$. In the study of Lipe & Salterio (2000) the *Division x Common* interaction effect is $\eta^2 = 0.353$, compared to $\eta^2 = 0.112$ in this current study. For the *Division x Unique* interaction effect the results of Lipe & Salterio (2000) show an eta-squared of 0.012, compared to $\eta^2 = 0.063$ in this current study. Thus, only 11.2% of the differences in performance evaluation scores can be explained by common measures, while 6.3% of the differences in performance evaluation scores can be explained by unique measures. These results suggest undergraduate students seem to rely more heavily on unique measures than M.B.A. students, while they rely less heavily on common measures in performance evaluation than M.B.A. students.

In figure 4.1 a graphical representation of the significant interaction effects found between *Division x Common* and *Division x Unique* is given. It shows the interaction of the common measures and unique measures of both divisions.

Figure 4.1. Graphical representation of interaction effects of common and unique measures with the within-subjects factor



An additional regression analysis, regressing the differences in performance evaluation, is performed to examine the relative influence of the common and unique measures on performance evaluation. In contrast to the findings of Lipe & Salterio (2000) who only reported a significant positive slope coefficient for the common measures, in this study both *common* and *unique* measures have significantly positive slope coefficients: 1.250 (t=4.781, p < 0.01) and 0.938 (t=3.588, p < 0.01) for *common* and *unique,* respectively. Thus, the results suggest that undergraduate students will use both common and unique measures in their performance evaluation, and there is no significant difference in the relative weighting of common and unique measures.

Table 4.3. Comparison of relative weights of common and unique measures on differences in subjective overall evaluations of division managers: regression analysis results (n = 56) [1]

Source	*df*	*Sum of squares*	*Mean square*	*F-value*	*p*
Model	2	34.134	17.067	17.846	0.000***
Residual	53	50.686	0.956		
Total	55	84.820			
R^2	0.402				
Adjusted R^2	0.380				

Variable	*df*	*Parameter estimate*	*Standard error*	*t-value*	*p*
Intercept	1	-3.210	0.577	-5.566	0.000***
Common	1	1.250	0.262	4.781	0.000***
Unique	1	0.938	0.262	3.588	0.001***

**P<0.10; ** P< 0.05; ***P<0.01*

[1] The dependent variable is the difference in the overall evaluations of RadWear's and WorkWear's division managers performance on a 13-point scale, with 0 labeled Reassign and 12 labeled Excellent.

Common = a 0/1 dummy variable indicating the particular division scored high (low) on the eight BSC measures that appeared on both divisions' scorecards (Roberts, et al., 2004).

Unique = a 0/1 dummy variable indicating the particular division scored high (low) on the eight BSC measures that were unique to that division, i.e. did not appear on both divisions' scorecards (Roberts, et al., 2004).

Concluded on this first hypothesis, it was found that common measure bias is not present in an experiment with undergraduate students. Undergraduate students use both common and unique measures in their performance evaluation of divisional managers, while M.B.A. students perceive the common measures as more important in performance evaluation. An analysis for eta-squared found that undergraduate students rely more on unique measures than M.B.A. students do.

Situation 2: Adjustment of mechanically aggregated scores

Secondly, it is tested whether adjustment of the mechanically aggregated score influences the results reported in table 4.1. The students who have to evaluate the managers on each separate measure of the BSC got the possibility to adjust their overall mechanically aggregated score. Only nine students have adjusted their mechanically aggregated judgment. So, it is expected that this does not cause differences in the results as presented in the first situation.

The results of this repeated measures ANOVA are given in table 4.4. It is found that both the *Division x Common* interaction effect (F = 22.947, p < 0.01) and the *Division x Unique* interaction effect (F = 12.806, p < 0.01) are significant. This means that also after adjustment of the mechanically aggregated scores, both common and unique measures are important in explaining differences in overall evaluation scores. Thus, also in this situation no common measure bias is found.

Table 4.4. Results of a 2 x 2 x 2 repeated measures ANOVA of evaluations of the performance of RadWear and WorkWear division managers (n = 56) [a]

Between-subjects	*df*	*Sum of squares*	*Mean Square*	*F*	*p*
Common	1	0.734	0.734	0.495	0.485
Unique	1	0.067	0.067	0.045	0.832
Common * Unique	1	5.449	5.449	3.678	0.061*
Error	52	77.043	1.482		

Within-subjects	*df*	*Sum of squares*	*Mean Square*	*F*	*p*
Division	1	0.327	0.327	0.645	0.426
Division * Common	1	11.626	11.626	22.947	0.000***
Division * Unique	1	6.488	6.488	12.806	0.001***
Division * Common * Unique	1	0.013	0.013	0.025	0.875
Error	52	26.345	0.507		

**P<0.10; ** P< 0.05; ***P<0.01*

[a] Evaluations made on a 13-point scale, with 0 labeled “Reassign” and 12 labeled “Excellent”

In conclusion to the first hypothesis, it was found that no common measure bias is found in an experiment with undergraduate students. Because this hypothesis is assumed to be the main hypothesis, the outcomes of the three sub-hypotheses will be used to decide whether the hypothesis is supported or not. On the basis of the above results this first hypothesis is apparently not supported. In the upcoming sections, the three sub hypotheses are tested which would indicate whether disaggregation of the balanced scorecard would attenuate common measure bias, and whether performance evaluation using a balanced scorecard affect compensation decisions.

H2: Disaggregation of performance evaluation will attenuate common measure bias in the balanced scorecard in an experiment with undergraduate students.

With respect to the second hypothesis, two different experimental conditions were analyzed to test whether disaggregation of the performance evaluation will attenuate common measure bias in the balanced scorecard.

First, the effect of disaggregation strategies on performance evaluation is tested. Table 4.5 shows the results of a repeated measure ANOVA for disaggregated judgments. It was found that when students were asked to assign sixteen separate judgments for every measure in the BSC, a significant interaction effect is found of *Division x Unique* (F=18.472, p<0.01) and that the interaction effect of *Division x Common* is marginally significant. This means that disaggregation of the BSC causes undergraduate students to pay more attention to unique measures, which are the measures that are tailored to the divisional strategy. So, they perceive the unique measures as more important in performance evaluation than the common measures.

Table 4.5. Results of a 2 x 2 x 2 repeated measures ANOVA of evaluations of the performance of RadWear and WorkWear division managers for disaggregated judgments (n = 27)

Between-subjects	*df*	*Sum of squares*	*Mean Square*	*F*	*p*
Common	1	0.368	0.368	0.237	0.631
Unique	1	1.864	1.864	1.199	0.285
Common * Unique	1	7.637	7.637	4.913	0.037
Error	23	35.750	1.554		

Within-subjects	*df*	*Sum of squares*	*Mean Square*	*F*	*P*
Division	1	1.663	1.663	5.538	0.028
Division * Common	1	1.974	1.974	6.571	0.017**
Division * Unique	1	5.548	5.548	18.472	0.000***
Division * Common * Unique	1	0.498	0.498	1.658	0.211
Error	23	6.908	0.300		

**P<0.10; ** P< 0.05; ***P<0.01*

[a] Evaluations made on a 13-point scale, with 0 labeled “Reassign” and 12 labeled “Excellent”

Second, there are also students who only have to assign an overall judgment to the divisional mangers. Table 4.6 shows that when students were asked to assign an overall judgment there is a significant interaction effect of *Division x Common* (F=23.658, p<0.01) and that the interaction effect of *Division x Unique* (F=3.452, p=0.075) is marginally significant. This indicates that when students were asked for an overall judgment, they pay more attention to the measures common to both divisions. Thus, while both common and unique measures affect the performance evaluation of undergraduate students, in this experimental condition a small degree of common measure bias was found.

Table 4.6. Results of a 2 x 2 x 2 repeated measures ANOVA of evaluations of the performance of RadWear and WorkWear division managers for overall judgment (n = 29)

Between-subjects	*df*	*Sum of squares*	*Mean Square*	*F*	*p*
Common	1	0.004	0.004	0.003	0.957
Unique	1	0.229	0.229	0.188	0.668
Common * Unique	1	0.916	0.916	0.753	0.394
Error	25	30.397	1.216		

Within-subjects	*df*	*Sum of squares*	*Mean Square*	*F*	*P*
Division	1	0.537	0.537	1.046	0.316
Division * Common	1	12.137	12.137	23.658	0.000***
Division * Unique	1	1.771	1.771	3.452	0.075*
Division * Common * Unique	1	0.537	0.537	1.046	0.316
Error	25	12.825	0.513		

**P<0.10; ** P< 0.05; ***P<0.01*

[a] Evaluations made on a 13-point scale, with 0 labeled "Reassign" and 12 labeled "Excellent"

In conclusion, the second hypothesis that *disaggregation of the performance evaluation will attenuate common measure bias in the balanced scorecard* is supported. The above results suggest that when students were asked to assign an overall judgment use more common measures in their performance evaluation, and students who receive an disaggregated balanced scorecard use more unique measures in their performance evaluation. However in both experimental conditions the common and unique measures affect the performance evaluation of undergraduate students.

This result differs from Roberts, Albright & Hibbets (2004), who found a significant interaction effect of both the common and the unique measures when the BSC is disaggregated. Thus, related to the main hypothesis, undergraduate students do not show the same degree of common measure bias as M.B.A. students when the BSC is disaggregated.

H3: Performance evaluations based on a balanced scorecard will affect subsequent bonus allocations in an experiment with undergraduate students.

The third hypothesis examines the influence of performance evaluation on bonus allocations of undergraduate students. The difference in bonuses of the two managers are calculated, which is used as the dependent variable of the regression analysis conducted. The differences in performance evaluation for the disaggregated and the overall performance evaluation were regressed. *DifferenceAggr* accounts for differences in mechanically aggregated scores and *DifferencePerf* accounts for differences in overall performance evaluation scores. The model applied is statistically significant (F=12.613, p<0.01).

Table 4.7. Influence of Disaggregated Balanced Scorecard Performance Evaluations on Difference in Managers' Bonuses: Regression Analysis Results (n=31) [1]

Source	*df*	*Sum of squares*	*Mean square*	*F-value*	*p*
Model	2	2,708,857,893	1,354,428,946	12.613	0.000***
Residual	28	3,006,831,535	107,386,840.5		
Total	30	5,715,689,428			
R^2	0.474				
Adjusted R^2	0.436				

Variable	*df*	*Parameter estimate*	*Standard error*	*t-value*	*p*
Intercept	1	2,736.574	1,962.603	1.394	0.174
DifferenceAggr	1	9,331.419	3,723.110	2.506	0.018**
DifferencePerf	1	8,115.276	1,909.984	4.249	0.000***

**P<0.10; ** P< 0.05; ***P<0.01*

[1] The dependent variable is the difference in dollar amounts of a total bonus of $100,000 that was available to allocate between the managers of RadWear and WorkWear.

It was found that the mechanically aggregated scores were marginally significant (t=2.506, p<0.018). The overall performance evaluation scores were found to be significant (t=4.249, p<0.01). This means that undergraduate students in both experimental conditions use performance evaluation scores in their bonus allocation decisions. Thus the third hypothesis, stating that performance evaluations based on a balanced scorecard will affect subsequent bonus allocations, is supported.

H4: Performance evaluations based on a balanced scorecard will affect subsequent promotion decisions in an experiment with undergraduate students.

The fourth hypothesis examines the influence of performance evaluation on promotion decisions of undergraduate students. By design, the students have to recommend promotion to either Chris Peeters (RadWear) or Bob Graham (WorkWear). This is the dependent variable of this regression analysis. Again, the differences in performance evaluation for the disaggregated and the overall performance evaluation were regressed. *DifferenceAggr* accounts for differences in mechanically aggregated scores and *DifferencePerf* accounts for differences in overall performance evaluation scores. The model applied is statistically significant (F=11.395, p<0.01).

Table 4.8. Influence of Disaggregated Balanced Scorecard Performance Evaluations on Promotion Decisions: Regression Analysis Results (n=25) [1]

Source	*df*	*Sum of squares*	*Mean square*	*F-value*	*p*
Model	2	2.564	1.282	11.395	0.000***
Residual	22	2.476	0.113		
Total	24	5.040			
R^2	0.509				
Adjusted R^2	0.464				

Variable	*df*	*Parameter estimate*	*Standard error*	*t-value*	*p*
Intercept	1	1.333	0.068	19.578	0.000
DifferenceAggr	1	-0.237	0.071	-3.357	0.003***
DifferencePerf	1	-0.236	0.069	-3.436	0.002***

**P<0.10; ** P< 0.05; ***P<0.01*

It was found that both the mechanically aggregated scores (t=-3.357, p<0.01) and the overall performance evaluation scores (t=-.3.436, p<0.01) were significant. This means that undergraduate students in both experimental conditions use performance evaluation scores in their promotion decisions. Thus the fourth hypothesis, stating performance evaluations based on a balanced scorecard will affect subsequent promotion decisions in an experiment with undergraduate students, is supported.

5 Conclusion

Common measure bias was defined as decision makers' unwillingness to incorporate the unique information because this information requires greater cognitive effort process (Hibbets, Roberts & Albright, 2004). Lipe & Salterio (2000) have used M.B.A. students to test whether common measure bias exists in the balanced scorecard. In this study their research is replicated with undergraduate students as experimental participants. Ultimately, the answering of the following research question was the primary goal of this research.

Does common measure bias in the balanced scorecard hold in an experiment with undergraduate students?

In the literature review the seven major experiments on attenuating common measure bias in the BSC were explored. This has led to the identification of five factors that possibly could attenuate common measure bias in the balanced scorecard: disaggregation of the balanced scorecard, linking bonus allocation to performance evaluation, training of participants, assurance reports and providing strategy information. The first two factors were chosen to investigate. This study contributes to the existing literature on these two factors in a way that Roberts, Albright & Hibbets (2004) did not incorporated disaggregation as a between-subjects factor. In their study all of the students receive the same experimental materials. Consequently, four hypotheses have been tested.

Hypothesis 2, stating that disaggregation of the balanced scorecard will attenuate common measure bias in the balanced scorecard, is supported. It was found that if the BSC is disaggregated undergraduate students pay more attention to the measures unique for a particular division than to the common measures. As opposed to the disaggregated, when students were asked for an overall judgment, more attention is paid to common measures than to the unique measures. So, the hypothesis is supported because the results suggest that disaggregation causes undergraduate students to use the unique measures in their performance evaluation.

Hypothesis 3, stating that performance evaluations based on a balanced scorecard will affect subsequent bonus allocations in an experiment with undergraduate students, is supported. It was found that undergraduate students use their performance evaluation to allocate bonuses to the divisional managers.

Hypothesis 4, stating that performance evaluations based on a balanced scorecard will affect subsequent promotion decisions in an experiment with undergraduate students, is supported. It was found that the promotion decisions of undergraduate students are made using the performance evaluation scores of the divisional managers.

Consequently, the main hypothesis, stating that common measure bias is also found in an experiment with undergraduate students, is not supported. The main hypothesis is not supported because it was found that common measure bias in the balanced scorecard does not hold in an experiment with undergraduate students. It was found that undergraduate students use both common and unique measures in their performance evaluation under all conditions tested. This is in contrast with the findings of Lipe & Salterio (2000) who found that only the common measures affect the performance evaluations of divisional managers. Table 5.1 provides a schematic overview of the hypotheses.

H1	Common measure bias is also found in an experiment with undergraduate students.	**Not supported**
H2	Disaggregation of performance evaluation will attenuate common measure bias in the balanced scorecard in an experiment with undergraduate students.	**Supported**
H3	Performance evaluations based on a balanced scorecard will affect subsequent bonus allocations in an experiment with undergraduate students.	**Supported**
H4	Performance evaluations based on a balanced scorecard will affect subsequent promotion decisions in an experiment with undergraduate students.	**Supported**

Table 5.1. Schematic overview of the hypotheses

Thus, as an answer on the main research question it could be stated that common measure bias *does not* hold in an experiment with undergraduate students. The results of this study suggest that the finding of Lipe & Salterio (2000) of common measure bias whereby common measures dominate unique measures does not hold in an experiment with undergraduate students. In contrast to their findings, in this study it was found that undergraduate students use both common and unique measures in performance evaluation under all conditions.

6 Discussion

This thesis reports the findings of an experiment investigating judgments made with the BSC by undergraduate students with a basic theoretical understanding of the BSC. The major finding is that undergraduate students use both common and unique measures in their performance evaluation and are consistent in their judgment across bonus allocation and performance evaluation. When the balanced scorecard is disaggregated they place a greater weight on unique measures, but also use common measures in their judgment. Thus, in all of the experimental situations there was no common measure bias found. In this chapter the interpretation of the results were given. Also the limitations of the study were discussed.

6.1. Interpretation of the results

This research extends the study of Lipe & Salterio (2000) but show entirely different results. There are two possible explanations for why the results from this study differ from that of Lipe & Salterio. First, the participants in the Lipe & Salterio study does not receive any training on the BSC prior to the experiment and it is therefore expected that the level of theoretical knowledge of the participants is lower than in this experiment. While no common measure bias is found in this experiment, the results of their study implies that decision makers place a greater weight on common measures than unique measures, as focusing on common measures is cognitively easier (Slovic & MacPhillamy, 1974; Lipe & Salterio, 2000). So, the results of this study may reflect the behavior of knowledgeable decision makers using a BSC, while the results from Lipe & Salterio reflect how decision makers initially use the BSC. A basic theoretical understanding of the balanced scorecard appeared to be more important than years of full time work experience in attenuating common measure bias.

Another factor that could cause the differences between the results of the two studies is the disaggregation of the balanced scorecard. Slovic & MacPhillamy (1974) argue that common measures dominated only in comparative judgments. As the disaggregation strategy asked for sixteen individual judgments, it is easier to differentiate judgment between the individual measures. The findings suggest that the students in this condition placed more emphasis on unique measures than on common measures, but use both in their performance evaluation. Thus, the differences in the results of this study may be due to a better theoretical understanding of the BSC, but also to the individual judgments of the experimental participants.

Furthermore, this study links performance evaluation to compensation decisions. The finding that the compensation decisions of undergraduate students are affected by their performance evaluation suggests that the BSC is useful as indicator for compensation decisions. Kaplan & Norton (2001) argue that compensation can be based up to 25 performance measures. Using the BSC as a base for compensation decision will heighten the interest of employees in all the performance measures of the BSC (Kaplan & Norton, 2001b). Thus, equal weighting of common and unique measure will enhance the usefulness of the BSC for purposes of compensation decisions.

6.2. Limitations

This study shows interesting and renewing results, but the experimental design deals with some limitations. First, although the experiment is carefully developed there are differences in the size of the experimental groups. This is caused by the voluntary participation in the experiment. On forehand 223 participants enrolled, of which 207 take part in the experiment. Although every experimental condition is fulfilled by at least one student, it is not possible to compare two experimental participants of exactly the same condition with each other for every experimental condition. Thus, a balanced design would enhance comparability across students. A second limitation is that the experiment is completed in one period. Although, consistent results were shown in this study, an experiment over multiple periods would strengthen the results of the study.

A third limitation of this study comprises the involvement of the students. Since the balanced scorecards of RadWear and WorkWear were adopted from Lipe & Salterio (2000) to enhance the comparability of the findings, the students were not involved in the development of the scorecards. However, these scorecards were used in the major experiments, and are therefore found to be useful in adequate testing of common measure bias in the BSC. In the experiment, the students take the role of the higher-level manager, who is in general involved in the development of the BSC. Thus, involvement of students in the development of the BSC would make it possible for students to understand the complete design of the BSC. Also, this would strengthen the findings on whether decision makers perceive the BSC the same as experimental participants with only a theoretical understanding of the BSC.

6.3. Implications of the results

6.3.1. Practical implications

The results of this research have practical implications in that it shows that when the decision makers of the organization really understand the concept and design of the BSC, it could be a useful tool for performance evaluation in an organization. The results of this research show that when decision makers have a theoretical understanding of the BSC, they use both the common and unique measures of the BSC in their performance evaluation. Lipe & Salterio (2000) state that underuse of unique measures reduces the potential benefits of the BSC because the unique measures are important in capturing the unit’s business strategy. Thus, using both the common and unique measures in performance evaluation will improve the understanding of the (divisional) targets.

Also, the subordinates have to understand the concept and design of the BSC to improve realization of the targets (performance measures) set in the BSC. In that case the BSC could be linked to compensation decisions (e.g. bonus allocation, career development).

6.3.2. Scientific implications

The evidence that common measure bias does not hold in an experiment with undergraduate students has two major scientific implications. First, all of the previous studies on common measure bias are conducted in North America. The differences in the results could be caused by the difference between European and American students. Further research should be conducted to test whether the use of different students would cause differences in the findings on common measure bias.

Second, not all of the undergraduate students included in the sample have a good understanding of the concept and design of the BSC. Only thirty-five (35) students answered at least nine of the twelve multiple-choice questions in the manipulation check right. A post-hoc analysis suggest that these 35 undergraduate students with apparently a thorough understanding of the BSC pay more attention to unique measures than to common measures. As can be seen in table 6.1, the *Division x Unique* interaction effect is significant (F=6.474, p=0.005) and the *Division x Common* interaction effect is only marginally significant (F=4.319, p=0.024).

Table 6.1. Results of a 2 x 2 x 2 repeated measures ANOVA of evaluations of the performance of RadWear and WorkWear division managers (n=35)

Between-subjects	*df*	*Sum of squares*	*Mean Square*	*F*	*p*
Common	2	9.947	4.973	2.892	0.173
Unique	2	4.272	2.136	1.242	0.305
Common * Unique	4	16.335	4.084	2.375	0.078*
Error	26	44.708	1.720		

Within-subjects	*df*	*Sum of squares*	*Mean Square*	*F*	*P*
Division	1	0.343	0.343	1.274	0.269
Division * Common	2	2.323	1.162	4.319	0.024**
Division * Unique	2	3.482	1.741	6.474	0.005***
Division * Common * Unique	4	2.468	0.617	2.294	0.086*
Error	26	6.993	0.269		

**P<0.10; ** P< 0.05; ***P<0.01*

Although the sample size of this post hoc analysis is small, it offers possibilities for future research. While no significant differences in common and unique measures were found in this study, this post hoc analysis shows a small degree of *unique measure bias*. Thus, students who are more knowledgeable about the concept and design of the BSC perceive unique measures slightly more important than the common measures.

6.4. Suggestions for future research

The major suggestion for future research is that undergraduate students could be used for research in managerial accounting. The results from this study show that a theoretical understanding of the concept and design of the BSC causes decision makers to pay more attention to the measures in a BSC that are unique to a particular division. In contrast to previous studies, these findings are renewing because no common measure bias was found in this study. With respect to future research, it could be suggested to use undergraduate students as (experimental) participants to conduct research in managerial accounting.

Also, other types of students could be used in research towards common measure bias in the BSC. For example, students studying for their master's degree could be used to verify the results found in this study. These students already achieved their bachelor's degree, but are expected to do not have relevant full-time work experience. This multi person approach could strengthen the results found in this study.

7 References

Babbie, E. (2010). *The Practice of Social Research*. Wadsworth: Cengage Learning.

Banker, D. R., Chang, H., & Pizzini, M. (2004). The Balanced Scorecard: Judgmental Effects of Performance Measures Linked to Strategy. *The Accounting Review, 79*(1), 1-23.

Banker, D. R., Chang, H., & Pizzini, M. (2011). The Judgmental Effects of Strategy Maps in Balanced Scorecard Performance Evaluations. *International Journal of Accounting Information Systems, 12*(4), 259-279.

Banker, R. D., Potter, G., & Srinivasan, D. (2000). An Empirical Investigation of an Incentive Plan that Includes Nonfinancial Performance Measures. *The Accounting Review, 75*(1), 65-92.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is Stronger than Good. *Review of General Psychology, 5*(4), 323-370.

Cardinaels, E., & Van Veen-Dirks, P. M. G. (2010). Financial Versus Non-Financial Information: The Impact of Information Organization and Presentation in a Balanced Scorecard. *Accounting, Organizations and Society, 35*(6), 565-578.

Cianci, A.M., Kaplan, S.E., & Samuels, J.A. (2013). The Moderating Effects of the Incentive System and Performance Measure on Managers' and Their Superiors' Expectations. *Behavioral Research in Accounting, 25*(1), 115-134.

Cheng, M. M., & Humphreys, K. A. (2012). The Differential Improvement Effects of the Strategy Map and Scorecard Perspectives on Managers' Strategic Judgments. *The Accounting Review, 87*(3), 899-924.

Dilla, W. N., & Steinbart, P. J. (2005). The Effects of Alternative Supplementary Display Formats on Balanced Scorecard Judgments. *International Journal of Accounting Information Systems, 6*(3), 159-176.

Dilla, W. N., & Steinbart, P. J. (2005a). Relative Weighting of Common and Unique Measures by Knowledgeable Decision Makers. *Behavioral Research in Accounting, 17*(1), 43-53.

DoHerty, E.M., & Nord, W.R. (1993). Compensation: Trends and Expanding Horizons. In R.T. Golembirwski (Ed.) *Handbook of Organizational Behavior.* Englewood Cliffs, NJ: Marcel Dekker.

Gagne, M. L., Hollister, J., & Tully, G. J. (2006). Using the Balanced Scorecard: Both Common and Unique Measures are Informative. *Journal of Applied Business Research, 22*(1), 147-160.

Gravetter, F.J., & Forzano, L.B. (2011). *Research Methods for the Behavioral Sciences.* Wadsworth: Cengage Learning.

Hibbets, A. R., Roberts, M. L., & Albright, T. L. (2006). *Common Measures Bias in the Balanced Scorecard: Cognitive Effort and General Problem Solving Ability*. Paper presented at the AAA 2007 Management Accounting Section (MAS) Meeting.

Humphreys, K. A., & Trotman, K. T. (2011). The Balanced Scorecard: The Effect of Strategy Information on Performance Evaluation Judgments. *Journal of Management Accounting Research, 23*(1), 81-98.

Ittner, C. D., Larcker, D. F., & Meyer, M. W. (2003). Subjectivity and the Weighting of Performance Measures: Evidence from a Balanced Scorecard. *The Accounting Review, 78*(3), 725-758.

Kang, G., & Fredin, A. (2012). The Balanced Scorecard: the Effects of Feedback on Performance Evaluation. *Management Research Review, 35*(7), 637-661.

Kaplan, R. S. (2010). *Conceptual Foundations of the Balanced Scorecard*. Working Paper. Harvard Business School. Retrieved from http://www.hbs.edu/faculty/Publication%20Files/10-074.pdf

Kaplan, R. S., & Norton, D. P. (1992). The Balanced Scorecard - Measures that Drive Performance. *Harvard Business Review, 70*(1), 71-79.

Kaplan, R. S., & Norton, D. P. (1996). Linking the Balanced Scorecard to Strategy. *California Management Review, 39*(1), 53-79.

Kaplan, R. S., & Norton, D. P. (2001a). Transforming the Balanced Scorecard from Performance Measurement to Strategic Management (Part I). *Accounting Horizons, 15*(1), 87-104.

Kaplan, R.S., & Norton, D.P. (2001b). Transforming the Balanced Scorecard from Performance Measurement to Strategic Management (Part II). *Accounting Horizons, 15* (2), 147-160

Kaplan, S.E., Petersen, M. J., & Samuels, J. A. (2012). An Examination of the Effect of Positive and Negative Performance on the Relative Weighting of Strategically and Non-Strategically Linked Balanced Scorecard Measures. *Behavioral Research in Accounting, 24*(2), 133-151.

Kaplan, S.E., & Wisner, P. S. (2009). The Judgmental Effects of Management Communications and a Fifth Balanced Scorecard Category on Performance Evaluation. *Behavioral Research in Accounting, 21*(2), 37-56.

Kaskey, V. L. (2008). *The Balanced Scorecard: A Comparative Study of Accounting Education and Experience on Common Measure Bias and Trust in a Balanced Scorecard*. Dissertation. School of Business and Technology. Capella University.

Kennedy, J. (1995). Debiasing the Curse of Knowledge in Audit Judgment. *The Accounting Review, 70*(2), 249-273.

Libby, R., & Libby, P.A. (1989). Expert Management and Mechanical Combination in Control Reliance Decisions. *The Accounting Review, 84*(4), 729-747.

Libby, T., Salterio, S. E., & Webb, A. (2004). The Balanced Scorecard: The Effects of Assurance and Process Accountability on Managerial Judgment. *The Accounting Review, 79*(4), 1075-1094.

Liedtka, S. L., Church, B. K., & Ray, M. R. (2008). Performance Variability, Ambiguity Intolerance, and Balanced Scorecard-Based Performance Assessments. *Behavioral Research in Accounting, 20*(2), 73-88.

Lipe, M. G., & Salterio, S. E. (2000). The Balanced Scorecard: Judgmental Effects of Common and Unique Performance Measures. *The Accounting Review, 75*(3), 283-298.

Lipe, M. G., & Salterio, S. E. (2002). A Note on the Judgmental Effects of the Balanced Scorecard's Information Organization. *Accounting, Organizations and Society, 27*(6), 531-540.

Lyness, K.S., & Cornelius, E.T. (1982). A Comparison of Holistic and Decomposed Judgment Strategies in a Performance Rating Simulation. *Organizational Behavior and Human Performance, 29*(1), 21-38.

Malina, M. A., & Selto, F. H. (2001). Communicating and Controlling Strategy: An Empirical Study of the Effectiveness of the Balanced Scorecard. *Journal of Management Accounting Research, 13*(1), 47-90.

Merchant, K. A., & Van der Stede, W. A. (2007). *Management Control Systems: Performance Measurement, Evaluation and Incentives*. Harlow: Pearson Educated Limited.

Neumann, B. R., Roberts, M. L., & Cauvin, E. (2010). Stakeholder Value Disclosures: Anchoring on Primacy and Importance of Financial and Nonfinancial Performance Measures. *Review of Managerial Science, 5*(2-3), 195-212.

Roberts, M. L., Albright, T. L., & Hibbets, A. R. (2004). Debiasing Balanced Scorecard Evaluations. *Behavioral Research in Accounting, 16*(1), 75-88.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

Slovic, P., & MacPhillamy, D. (1974). Dimensional Commensurability and Cue Utilization in Comparative Judgment. *Organizational Behavior and Human Performance, 11*(2), 172-194.

Sprinkle, G.B. (2003) Perspectives on experimental research in managerial accounting. *Accounting, Organizations and Society, 28* (1)*,* 287-318.

Tayler, W. B. (2010). The Balanced Scorecard as a Strategy-Evaluation Tool: The Effects of Implementation Involvement and a Causal-Chain Focus. *The Accounting Review, 85*(3), 1095-1117.

Ullrich, M.J., & Tuttle, B.M. (2004). The Effects of Comprehensive Information Reporting Systems and Economic Incentives on Managers' Time-Planning Decisions. *Behavioral Research in Accounting. 16*(1), 89-105.

Wong-on-Wing, B., Guo, L., Li, W., & Yang, D. (2007). Reducing Conflict in Balanced Scorecard Evaluations. *Accounting, Organizations and Society, 32*(4-5), 363-377.

APPENDIX I: Balanced Scorecard

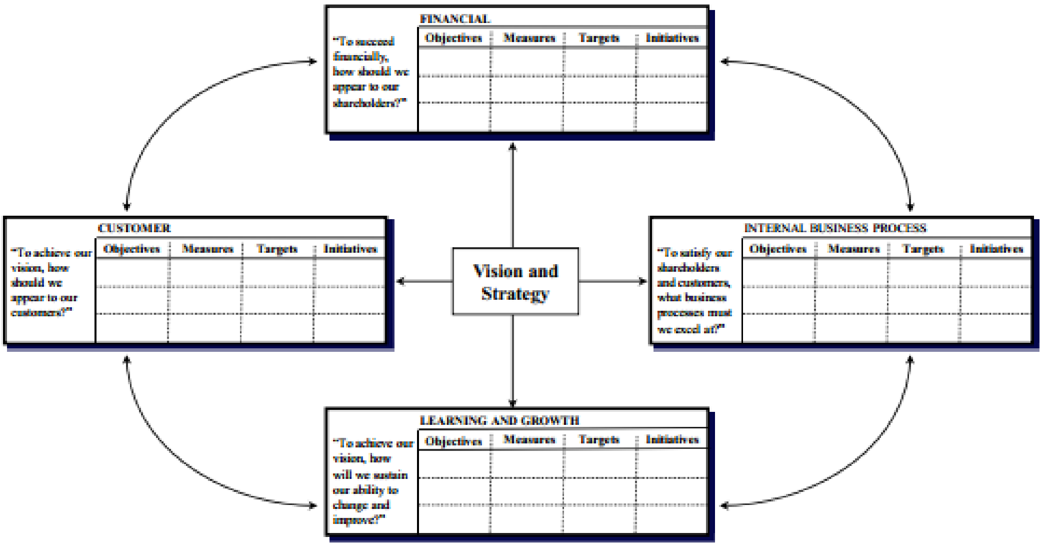

FINANCIAL
"To succeed financially, how should we appear to our shareholders?"
Objectives
Measures
Targets
Initiatives
CUSTOMER
"To achieve our vision, how should we appear to our customers?"
Objectives
Measures
Targets
Initiatives
Vision and Strategy
INTERNAL BUSINESS PROCESS
"To satisfy our shareholders and customers, what business processes must we excel at?"
Objectives
Measures
Targets
Initiatives
LEARNING AND GROWTH
"To achieve our vision, how will we sustain our ability to change and improve?"
Objectives
Measures
Targets
Initiatives

Figure 1: The Balanced Scorecard, source: Kaplan (2010)

APPENDIX II: Experimental materials

Instructions:

- During the experiment, which is part of the closed book exam for MAC,
  - I have my student card with me
  - I shall be seated at the place, which is Block: , Row: , Seat: , that is assigned to me
  - I shall only talk to the supervisors. If some wordings in the case are not clear to you, ask the supervisors for translations. Also if you are unsure about how to fill in the evaluation forms contact the supervisors for assistance. In any case it is forbidden to talk to other students during the experiment.
  - I shall only use pen (black or blue ink), and optionally a simple (non-programmable) calculator and a dictionary (but no electronic dictionary) for translating English words to my native tongue.
  - No food and drinks are allowed
  - The used of phones, mail, Internet or electronic communication is not allowed.

- After reading the information on the instructions and time reservations, you should start by reading the case material carefully.

  Bear in mind that you will be asked to take the role of a senior executive of Chadwick Limited. This executive is the direct boss of Chris Peeters and Bob Graham. He reviews Chris's and Bob's performance at the end of the year. The results of year-end reviews are used in determining Chris and Bob's merit raises and year-end bonuses.

- After you feel comfortable with the case material, go to the division scorecards to assess the performance of Chris and Bob.
- The complete the evaluation forms.
- Then complete Exhibit 3.
- Before moving on, put the first part of the form in the envelope. The first part contains the instruction page, the case material and the pages with Exhibit 1 and Exhibit 2. Seal the envelope and put your name on it.
- Finally complete the debriefing questionnaire. This is part 2.
- You only get grades for the exam if you have filled in **all sections of the experiment (part 1 and 2).**

Time

There is no strict time limit to fill in this form. In principle, take the time you think that is needed. The students assistants are reserved from 15h45 till 17h30. For students who need more time, Tom De Schryver will keep you company till 19h. If you leave your seat after 17h30, you can no longer return to it. Hence, if you need to go to the toilets do it before 17h30.

Note that you only get grades for the exam if you have filled in all sections of the experiment. So, take the time that you think is necessary.

APPENDIX III: BSCs of RadWear and WorkWear

	Target	Actual	%better than target
Financial			
Return on sales	24%	25%	4,17%
New store sales	30%	33%	8,33%
Sales growth	35%	37%	5,88%
Market share relative to retail space	£ 80	£ 86,55	8,56%
Customer			
Mystery shopper program rating	85	96,0	12,94%
Repeat sales	30%	32%	8,00%
Returns by customers as % of sales	12%	11,6%	-3,33%
Customer satisfaction rating	92%	94%	2,38%
Internal business			
Returns to suppliers	6%	5%	-12,50%
Average major brand names/store	32	37,0	15,63%
Average markdowns	16%	14,8%	-7,50%
Sales from new market leaders	25%	26%	16,00%
Learning & Growth			
Average tenure of sales personnel	1,4	1,6	14,29%
Hours of employee training/employee	15	16	6,67%
Stores computerizing	85%	90,0%	5,88%
Employee suggestions / employee	3,3	3,4	3,03%

	Target	Actual	%better than target
Financial			
Return on sales	24%	25%	4,17%
Revenues per sales visit	£ 400	£ 433,33	8,33%
Sales growth	34%	36%	5,88%
Online sales	6%	6,5%	8,56%
Customer			
Captured customers	20%	22,6%	12,94%
Repeat sales	25%	27%	8,00%
Referrals	50%	51,6%	3,20%
Customer satisfaction rating	84%	86%	2,38%
Internal business			
Returns to suppliers	8%	7,0%	-12,50%
Orders filled within one week	85%	99,0%	16,47%
Average markdowns	20%	18,5%	7,50%
Web orders filled with errors	5%	4,2%	16,00%
Learning & Growth			
%staff with M.B.A degrees	12%	13,6%	13,33%
Hours of employee training/employee	12	13,0	8,33%
Database certification of clerks	20%	21,2%	6,00%
Employee suggestions / employee	3,1	3,2	3,23%

APPENDIX IV: Manipulation check

For each statement, put ONE mark (X) to indicate to what extent you agree with the statements	Strongly disagree					Neither agree nor disagree					Strongly agree
	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
The two divisions, RadWear and WorkWear, use different performance measures											
The two divisions, RadWear and WorkWear, sell to the same markets											
It was appropriate for RadWear and WorkWear to employ different performance measures											
The strategy of RadWear is to generate greater sales through its existing infrastructure rather than invest in new stores											
The strategy of WorkWear is to target the companies more than the employees											
To grow ales, RadWear must successfully introduce new lines of clothing to its existing customers.											
WorkWear relies on new distibution channels to retain existing customers.											
The case material was easy to understand.											
The case was easy to do.											
The case material was realistic.											

APPENDIX V: Debriefing questionnaire

Questions related to you personally:

- Did you visit a retail-clothing store in the last 12 months (like Mexx, Esprit, C&A…)? NO/YES

- Did you ever work in a retail-clothing store (like Mexx, Esprit, C&A…)? NO/YES.

  If you answer YES how long did you work in total? (Multiple answers are unlikely, but possible)

  □ Less than a month or during one school holiday
  □ During multiple school holidays
  □ Part time during a school year
  □ Part time during multiple school years
  □ Full time for at least 3 months but no more than a year
  □ Full time for more than a year
  □ Other: please specify____

- You are MALE/FEMALE.

- For this course, you are enrolled as a
  □ BIT student
  □ BK student
  □ BSK student
  □ GZW student
  □ TBK student
  □ Other student. Please specify which study program:____

- Beyond this course (Management Accounting and Control), how many other courses have you followed in which the Balanced scorecard has been thought? The number of relevant courses is _____.

  How old are you? (in years) _____________

- How many years of FULL time working experience do you have? (Write NONE if you do not have FULL time working experience). My FULL TIME work experience is__________(years).

- If you have FULL time work experience was it in
  □ accounting, auditing, or taxation
  □ marketing or sales
  □ the retail industry

APPENDIX VI: Supplemental analysis

In this section, the analyses of chapter 4 were repeated for all experimental participants. Whereas the previous analysis only include the experimental conditions in which RadWear or WorkWear is favored. The analyses below include also the experimental conditions in which no division is favored. Thus, all 207 experimental participants are included. Consequently, the degrees of freedom (df) are higher in these analyses because now there are more situations:

- Favor RadWear on common / unique measures
- Favor WorkWear on common / unique measures
- No division favored

Thus, the analyses as reported in the fourth chapter were repeated for all the experimental participants. No differences were found between the findings of chapter 4 and the findings in these analyses. This means that also when no division is favored no common measure bias is found under all conditions.

Table 1: Results of a 2 x 2 x 2 repeated measures ANOVA of evaluations of the performance of RadWear and WorkWear division managers

Between-subjects	*df*	*Sum of squares*	*Mean Square*	*F*	*p*
Common	2	0.473	0.237	0.117	0.890
Unique	2	1.573	0.787	0.388	0.679
Common * Unique	4	13.658	3.415	1.685	0.155
Error	198	401.139	2.026		

Within-subjects	*df*	*Sum of squares*	*Mean Square*	*F*	*P*
Division	1	0.034	0.034	0.084	0.772
Division * Common	2	17.551	8.776	21.978	0.000***
Division * Unique	2	18.716	9.358	23.437	0.000***
Division * Common * Unique	4	3.760	0.940	2.354	0.055*
Error	198	79.058	0.399		

**P<0.10; ** P< 0.05; ***P<0.01*

Table 3: Comparison of relative weights of common and unique measures on differences in subjective overall evaluations of division managers: regression analysis results (n=207)

Source	*df*	*Sum of squares*	*Mean square*	*F-value*	*p*
Model	2	86.742	43.371	53.224	0.000***
Residual	204	166.236	0.815		
Total	206	252.978			
R^2	0.343				
Adjusted R^2	0.336				

Variable	*df*	*Parameter estimate*	*Standard error*	*t-value*	*p*
Intercept	1	0.020	0.063	0.325	0.745
DifferenceAggr	1	-0.586	0.088	-6.645	0.000***
DifferencePerf	1	-0.685	0.087	-7.891	0.000***

**P<0.10; ** P< 0.05; ***P<0.01*

Table 4: Results of a 2 x 2 x 2 repeated measures ANOVA of evaluation of the performance of RadWear and WorkWear division managers (n=207)

Between-subjects	*df*	*Sum of squares*	*Mean Square*	*F*	*p*
Common	2	0.338	0.169	0.088	0.916
Unique	2	1.352	0.676	0.351	0.704
Common * Unique	4	12.599	3.150	1.638	0.166
Error	198	380.735	1.923		

Within-subjects	*df*	*Sum of squares*	*Mean Square*	*F*	*P*
Division	1	0.021	0.021	0.051	0.822
Division * Common	2	18.842	9.421	22.379	0.000***
Division * Unique	2	18.842	9.421	22.379	0.000***
Division * Common * Unique	4	3.814	0.954	2.265	0.064
Error	198	83.352	0.421		

**P<0.10; ** P< 0.05; ***P<0.01*

Table 5: Results of a 2 x 2 x 2 repeated measures ANOVA of evaluations of the performance of RadWear and WorkWear division managers for disaggregated judgments (n=101) [a]

Between-subjects	*df*	*Sum of squares*	*Mean Square*	*F*	*p*
Common	2	0.127	0.063	0.037	0.964
Unique	2	4.375	2.188	1.278	0.283
Common * Unique	4	10.236	2.559	1.495	0.210
Error	92	157.469	1.712		

Within-subjects	*df*	*Sum of squares*	*Mean Square*	*F*	*P*
Division	1	1.252	1.252	9.794	0.002***
Division * Common	2	4.417	2.209	17.273	0.000***
Division * Unique	2	11.040	5.520	43.170	0.000***
Division * Common * Unique	4	1.119	0.280	2.188	0.076*
Error	92	79.058	0.399		

**P<0.10; ** P< 0.05; ***P<0.01*

Table 6: Results of a 2 x 2 x 2 repeated measures ANOVA of evaluation of the performance of RadWear and WorkWear division managers for overall judgment (n=106)

Between-subjects	*df*	*Sum of squares*	*Mean Square*	*F*	*p*
Common	2	1.736	0.868	0.389	0.679
Unique	2	1.489	0.744	0.334	0.717
Common * Unique	4	7.096	1.774	0.795	0.531
Error	97	216.428	2.231		

Within-subjects	*df*	*Sum of squares*	*Mean Square*	*F*	*P*
Division	1	0.734	0.734	1.192	0.278
Division * Common	2	16.611	8.306	13.479	0.000***
Division * Unique	2	9.654	4.827	7.834	0.001***
Division * Common * Unique	4	5.612	1.403	2.277	0.066*
Error	97	59.772	0.616		

**P<0.10; ** P< 0.05; ***P<0.01*

Table 7: Influence of Disaggregated Balanced Scorecard Performance Evaluation on Differences in Managers' Bonuses: Regression Analysis Results (n=117)

Source	*df*	*Sum of squares*	*Mean square*	*F-value*	*p*
Model	2	9,327,263,894	4,663,631,947	22.416	0.000***
Residual	114	24,006,936,523	210,587,162.5		
Total	116	33,334,200,417			
R^2	0.280				
Adjusted R^2	0.267				

Variable	*df*	*Parameter estimate*	*Standard error*	*t-value*	*p*
Intercept	1	-591.480	1,350.614	-0.438	0.662
DifferenceAggr	1	10,440.386	2,716.271	3.844	0.000***
DifferencePerf	1	7,731.137	1,422.982	5.433	0.000***

**P<0.10; ** P< 0.05; ***P<0.01*

Table 8: Influence of Disaggregated Balanced Scorecard Performance Evaluations on Promotion Decisions: Regression Analysis Results (n=90)

Source	*df*	*Sum of squares*	*Mean square*	*F-value*	*p*
Model	2	7.659	3.930	21.181	0.000***
Residual	87	15.730	0.181		
Total	89	23.389			
R^2	0.327				
Adjusted R^2	0.312				

Variable	*df*	*Parameter estimate*	*Standard error*	*t-value*	*p*
Intercept	1	1.389	0.045	30.899	0.000***
DifferenceAggr	1	-0.358	0.076	-4.695	0.000***
DifferencePerf	1	-0.203	0.045	-4.494	0.000***

**P<0.10; ** P< 0.05; ***P<0.01*