Supporting Material is available for this work. For more information, follow the link from the Table of Contents to "Accessing Supporting Material".

Simulation Based Experiments for Testing the Balanced Scorecard's Built-in Performance Improvement Theory

Jürgen Strohhecker

HfB / Business School of Finance and Management
Sonnemannstraße 9-11
D-60314 Frankfurt am Main
Germany

Telephone: +49 69 154008-110
E-Mail: strohhecker@hfb.de

Kaplan and Norton's balanced scorecard (BSC) is, without a doubt, one of the last decade's major improvements in management and controlling tools. In their various articles and books, the authors maintain that use of a balanced scorecard will ultimately improve an organisation's performance. Their theory about the scorecard's performance impact, however, is not explicitly described. Based on Kaplan and Norton's publications, this article reconstructs a system of hypotheses about the impact of a balanced scorecard on performance and describes a research design, which uses a System Dynamics-based micro-world, to test the theory. The implementation of the research design is portrayed and statistically tested. Finally, some preliminary results are presented, indicating that the balanced scorecard's effect on organisational performance might be overestimated.

Key Words: Balanced Scorecard; Empirical Research; Research Design; Simulation Experiment

1 The Balanced Scorecard Success Story

Kaplan and Norton's 1992 Harvard Business Review article "The Balanced Scorecard – Measures that Drive Performance" has triggered an avalanche of publications and projects. Since then, dozens of books about the balanced scorecard have been published, and hundreds of articles have been written. The number of organisations actually using a balanced scorecard has risen steadily although the true degree of adoption is difficult to estimate. Reliable empirical research in this area is still scarce. However, the studies available hint at a fairly rapid adoption process, indicating that the balanced scorecard is meeting a management need:

- In his 1999 survey of management tools and techniques, Rigby (2001, 143) found that 43.9 % of all responding North American companies were using the balanced scorecard.
- In Germany, 46 % of the top 200 companies asked during spring and summer 2000 indicated that they were using the balanced scorecard (PWC Deutsche Revision 2001).

Even more astonishing than the fast adoption process is the lack of warning or critical voices among scholars. Atkinson et. al. (1997, 94), for example, consider the balanced scorecard "among the most significant developments in management accounting" and Mooraj et. al. (1999, 489) conclude that the balanced scorecard "is a 'necessary good'

for today's organisations". Many articles can be found describing the features of the balanced scorecard and its development and implementation process in different types of industries and organisations (e.g. Kaplan/Norton 1992, 1993, 1996a, Goulian/Mersereau 2000). Most of these articles also contain cases and examples to illustrate the application of the concept. The authors' final judgement is for the most part positive: the overwhelming majority of the papers report success; very few admit failures (e.g. Ho/McKay 2002). Critical examinations of the balanced scorecard concept are rarely found in business publications. One exception is Nørreklit's paper in Management Accounting Research (2000). If there is mention of a limiting factor at all, it is the warning that the implementation of a company-wide balanced scorecard system is usually more time consuming and costly than it would appear at a first glance.

While scholars seldom criticising the balanced scorecard, organisations already using the tool are mostly satisfied. Apart from a significant number of case studies, which report the successful development and introduction of a balanced scorecard, there is some additional evidence for the satisfaction with the balanced scorecard in the surveys quoted above. In Rigby's study (2001, 145) the balanced scorecard attained an average of 3.85 on a scale of 1 (dissatisfied) to 5 (extremely satisfied). This score is just slightly above the overall average of 3.76, but ranked as eighth of 25 management tools. In the German survey no overall satisfaction was determined; instead, the companies had the possibility to check multiple criteria expressing their perceived utility. Despite the general impression of high satisfaction among German companies, the lack of an appropriately scaled question makes an accurate and just judgement impossible.

Why is the balanced scorecard flourishing? It seems to have specific characteristics that help to successfully address some of the problems traditional tools for managing and controlling businesses cannot not solve.

2 The Balanced Scorecard's Characteristics and Its Stated Benefits

In their 1992 article, Kaplan and Norton introduced the balanced scorecard mainly as a balanced performance measurement system with a comprehensible number of indicators.[1] The importance of a company's overall vision and strategy for the development process of a balanced scorecard was seen, yet the implications on the strategic management process were not worked out. Subsequent books and articles – (Kaplan/Norton 1996a + b, Kaplan/Norton 2001) – have placed increasing emphasis on the connection between measurement and strategy and have extended the balanced scorecard to a strategic management system. Taking all the published articles and books into account, it seems to be justified to regard the balanced scorecard as an operational and strategic management system with the following core components:

1. The balanced scorecard report. The report is drawn up regularly, e.g. monthly, and provides an overview of the actual outcome of a set of about 10 to 25 measures, each compared to the target and historical values. The measures are related to four important perspectives – (1) financial, (2) customer, (3) internal business, and (4)

[1] Kaplan and Norton (1996b) recommend between 4 and 7 measures per perspective and between 16 and 25 measures for the whole scorecard.

innovation and learning – ensuring a broad and balanced information feedback. See Figure 1 for an example of a balanced scorecard report.

2. The balanced scorecard development methodology. In their articles and books, Kaplan and Norton prescribe a set of rules and recommendations how to develop the balanced scorecard metrics in order to derive the greatest possible utility. One such recommendation is, for example, the important advice to align the balanced scorecard metrics with the organisation's vision and strategy and to aim for a multidimensional balanced set of measures (financial and non-financial, leading and lagging, tangible and intangible).
3. The causal hypothesis system. In order to develop a balanced scorecard in alignment with the organisation's strategy, the strategy has to be expressed as a system of causal hypotheses. The causal hypotheses system is therefore the backbone of a strategy-aligned balanced scorecard. Causal hypotheses can be formulated as if-then-sentences and can also be visually represented. Kaplan and Norton use different diagramming techniques to visualise the cause-and-effect model that forms the organisation's strategy: the latest invention is the strategy map (see Figure 1 for an example), but they also depict strategy tress (2001, 228) and different types of word-and-arrow diagrams (1996a, 83).
4. The balanced scorecard management system methodology. Like the development methodology, the management system methodology consists of a set of rules and recommendations. It tells the user how to apply the balanced scorecard in order to build an operational and strategic management system, or in Kaplan and Norton's own words, how to build "the strategy focused organisation".

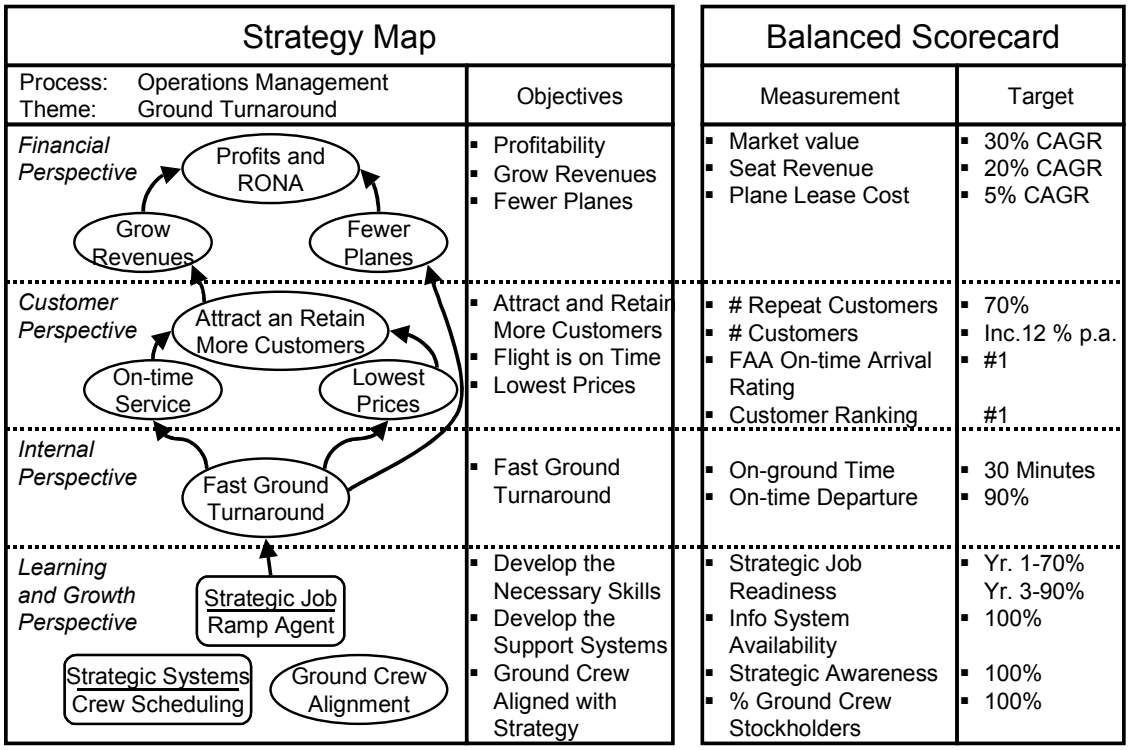

Figure 1: Strategy Map and Balanced Scorecard (Kaplan/Norton 2004, 44)

In their books and articles Kaplan and Norton emphasise significant benefits associated with using the balanced scorecard. From today's perspective the most important statements concerning derived benefits are the following:

- The balanced scorecard "gives top managers a fast but comprehensive view of the business" (Kaplan/Norton 1992, 71).
- "While giving senior managers information from four different perspectives, the balanced scorecard minimizes information overload by limiting the number of measures used" to 16 – 25 (Kaplan/Norton 1992, 72).
- The "scorecard guards against suboptimization. By forcing senior managers to consider all the important operational measures together, the balanced scorecard lets them see whether improvement in one area may have been achieved at the expense of another" (Kaplan/Norton 1992, 73)
- "The scorecard puts strategy and vision, not control, at the centre" (Kaplan/Norton 1992, 79).
- "The balanced scorecard can serve as a focal point for the organization's efforts, defining and communicating priorities to managers, employees, investors, even customers "(Kaplan/Norton 1993, 135)
- "The balanced scorecard enables a company to align its management processes and focuses the entire organization on implementing long term strategy." It links "a company's long-term strategy with its short-term actions" (Kaplan/Norton 1996a, 85, 73)
- "The very exercise of creating a balanced scorecard forces companies to integrate their strategic planning and budgeting processes and therefore helps to ensure that their budgets support their strategies" (Kaplan/Norton 1996a, 82).
- The "scorecard facilitates the strategy review that is essential to strategic learning" what means that the scorecard supports "what Chris Argyris calls double-loop learning – learning that produces a change in people's assumptions and theories about cause-and-effect relationships" (Kaplan/Norton 1996a, 85, 84)
- "Breakthrough results = {Strategy Maps} + {Balanced Scorecard} + {Strategy-Focused Organization}" (Kaplan/Norton 2004b, xiii).

Essentially, the balanced scorecard promises support for both operational or day-to-day management as well as for strategic management. Kaplan and Norton believe that using a balanced scorecard will increase both long-term and short term performance. In the following chapter, the balanced scorecard's built-in theory about its impact on business performance will be elaborated in more detail.

3 The Balanced Scorecard's Built-In Theory about Its Impact on Business Performance

Kaplan and Norton do not explicitly and systematically describe the balanced scorecard's built-in theory about its impact on business performance. However, as the statements listed above illustrate, Kaplan and Norton clearly suggest a balanced scorecard will, in the end, improve business performance. The various means and arguments they discuss indicate that they see a broad range of cause-and-effect relationships. For a more rigorous analysis and for empirical testing, however, it is

necessary to condense their articles and books into a system of hypotheses concerning the balanced scorecard's impact on performance.

Kaplan and Norton see management's decision-making and problem-solving competence as major driver of an organisation's performance (Kaplan/Norton 1992, 79). Improving management's decision-making and problem-solving capabilities will therefore eventually lead to better performing organisations. For Kaplan and Norton the balanced scorecard is obviously the means to improve a manager's decision-making ability. The global performance impact hypothesis can therefore be formulated as follows:

H_1: If the management of an organisation uses a balanced scorecard as management and controlling system, the organisation's performance will increase.

Better decision-making is the result of an improved learning process. Kaplan and Norton mention explicitly enhanced double loop learning (Kaplan/Norton 1996a, 84). But there is also a single loop learning process that can be improved (Argyris 1976, 365, 367-368). While the double loop learning process relates to strategic management, single loop learning can be regarded as the domain of operational management.

Single loop learning can be understood as a goal-seeking feedback control loop (Figure 2): Managers compare the actual situation with the desired one and – guided by the organisation's strategy – do something if there is a deviation (Forrester 1961, Argyris 1976, Kim 1993, Argyris 1999). Normally the intensity of the action rises as the deviation grows, moving the actual situation towards the goal and narrowing the gap. Eventually, if no external forces change the actual situation or the aimed position exogenously, the gap is closed and the single loop learning process has successfully moved the organisation to the new equilibrium state.

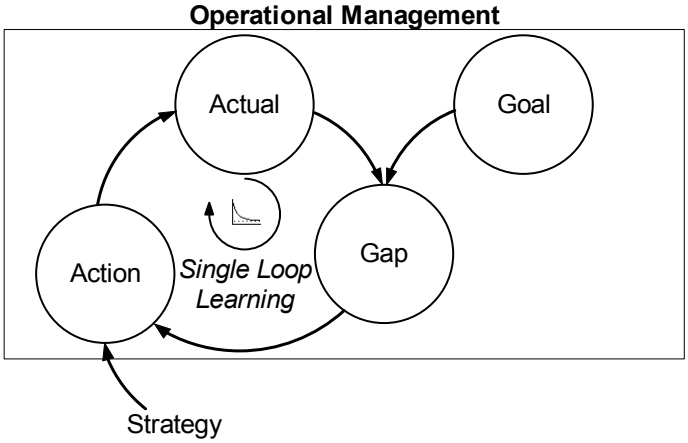

Figure 2: Single Loop Learning and Operational Management (Strohhecker 2002, 9)

"Double loop learning occurs when mismatches are corrected by first examining and altering the governing variables and then the actions" (Argyris 1999, 68). For strategic management, double loop learning implies to question and – if necessary – to change an organisation's strategy (Figure 3). The strategic management process, which includes strategic analysis, strategy development, strategy evaluation and implementation, is therefore part of the double loop learning process.

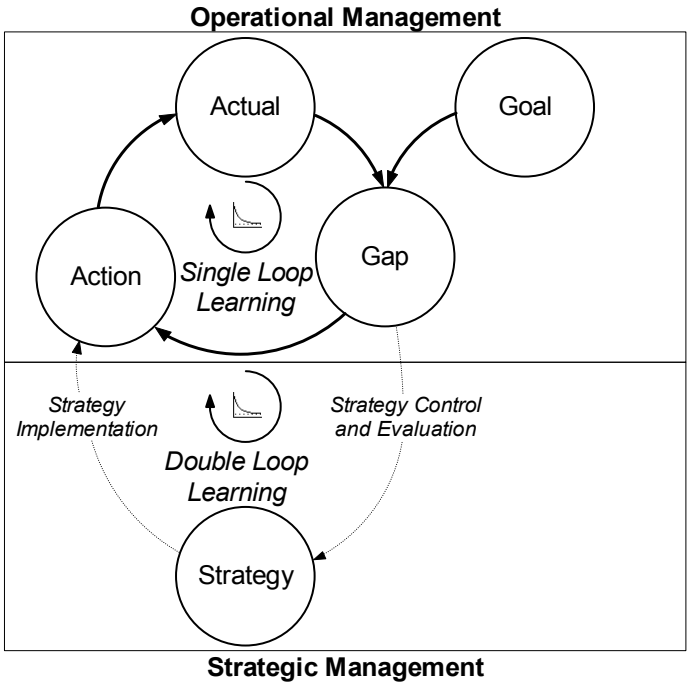

Figure 3: Double Loop Learning and Strategic Management

Since the balanced scorecard is supposed to improve both single and double loop learning, two more detailed hypotheses can be formulated:

H_{2a}: If an organisation's single and double loop learning is accelerated, its performance will increase.

H_{2b}: If the management of an organisation uses a balanced scorecard as management and controlling system, the organisation's single and double loop learning will be accelerated.

H_{2a} and H_{2b} are equivalent to hypothesis H_2:

H_2: If the management of an organisation uses a balanced scorecard as management and controlling system, the organisation's single and double loop learning is accelerated, which will increase its performance.

In order to further elaborate the theory, single loop learning and its drivers are more closely examined in a first step. According to Sterman, enabling bounded rational human actors to make good decisions in the single loop learning context requires especially fast delivery of the right information without causing overload (Sterman 2000, p. 14-33). This means that there are three causes for successful single loop learning: (a) information delivery lead-time, (b) information overload, and (c) information quality. As a result, three further hypotheses can be formulated:

H_4: If the management of an organisation uses a balanced scorecard as management and controlling system, information delivery lead-time in the organisation is reduced, which accelerates single loop learning which in turn leads to increased performance.

H_5: If the management of an organisation uses a balanced scorecard as management and controlling system, the information overload of the organisation's managers is reduced, which accelerates single loop learning and leads to increased performance.

H_6: If the management of an organisation uses a balanced scorecard as management and controlling system, the information quality in the organisation improves, which accelerates single loop learning and leads to increased performance.

In academic publications, information quality is seen as multidimensional concept (Miller 1996, Wang/Strong 1996, Klein 2001). Following the user-focused approach of Wang and Strong (1996), four drivers of information quality can be distinguished: (1) accuracy, (2) accessibility, (3) representation, and (4) relevance of the information provided. While accuracy indicates the extent to which data values are in conformance with the actual or true values, and accessibility specifies the availability of the information, representation measures the extent to which the information is presented in an intelligible and clear manner. Finally, relevance covers the extent to which the data are applicable (pertinent) to the task of the user.

Kaplan and Norton do not claim data accuracy as a specific advantage offered by the balanced scorecard. Therefore no hypothesis is included for this factor, leaving three further hypotheses:

H_7: If the management of an organisation uses a balanced scorecard as management and controlling system, information accessibility is simplified, which improves information quality, which in turn accelerates single loop learning and leads to increased performance.

H_8: If the management of an organisation uses a balanced scorecard as management and controlling system, information representation gets better, which improves information quality, which in turn accelerates single loop learning and leads to increased performance.

H_9: If the management of an organisation uses a balanced scorecard as management and controlling system, the relevance of the information is increased, which improves information quality, which in turn accelerates single loop learning and leads to increased performance.

Having finished the hypotheses on single loop learning, the hypotheses covering the causes of improved double loop learning are developed. In accordance with Figure 3, two mechanisms can be distinguished, resulting in the formulation of hypotheses 10 and 11:

H_{10}: If the management of an organisation uses a balanced scorecard as management and controlling system, strategy evaluation and control is improved which accelerates double loop learning and leads to increased performance.

H_{11}: If the management of an organisation uses a balanced scorecard as management and controlling system, strategy implementation is improved which accelerates double loop learning and leads to increased performance.

Kaplan and Norton see the strategic management benefit resulting from the balanced scorecard not only in improved double loop learning but also in better strategy implementation and better strategy control and evaluation. These two additional cause-and-effect relationships are shown in Figure 4.

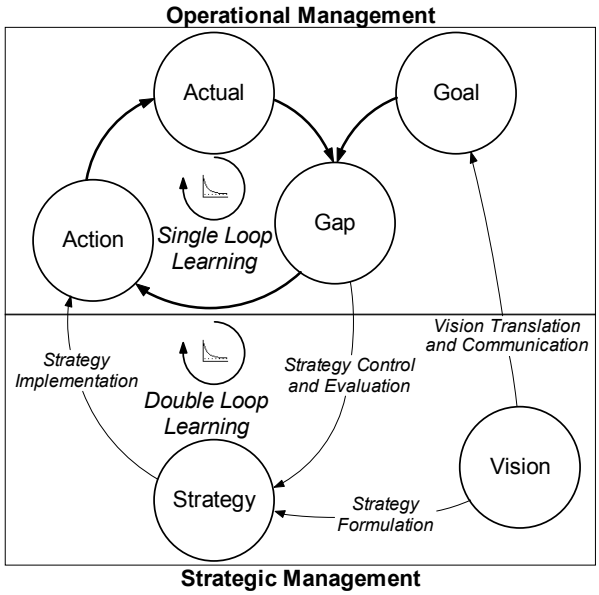

Figure 4: Vision Dissemination

Hence, hypotheses 12 and 13 can be formulated as follows:

H_{12}: If the management of an organisation uses a balanced scorecard as management and controlling system, the translation of an organisation's vision into concrete and measurable goals and its communication is improved which accelerates double loop learning and leads to increased performance.

H_{13}: If the management of an organisation uses a balanced scorecard as management and controlling system, the translation of an organisation's vision into the strategy is improved which accelerates double loop learning and leads to increased performance.

The complete hypotheses system about the balanced scorecard's impact on performance is graphically shown in Figure 5.

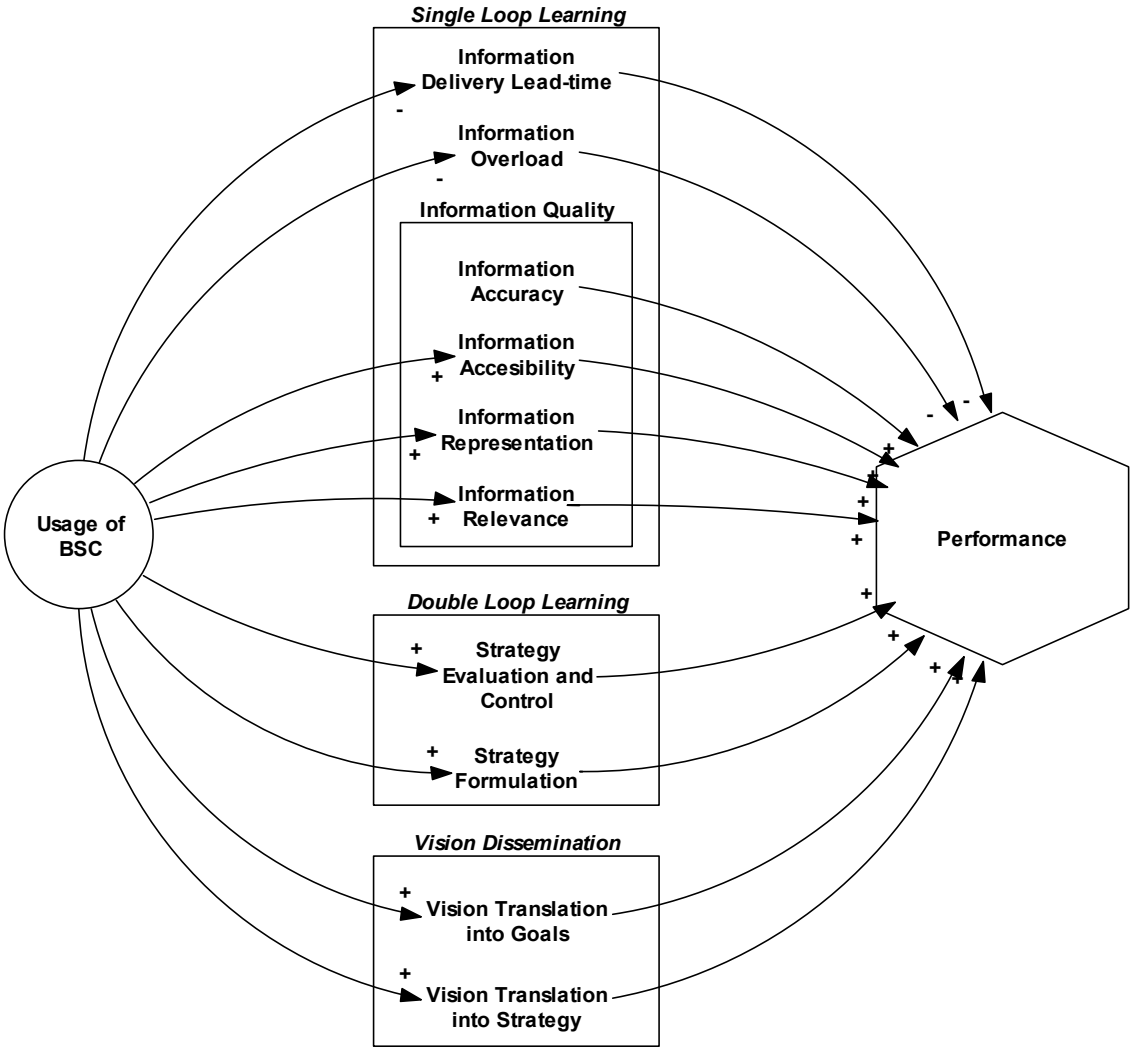

Figure 5: Hypotheses System on BSC's Impact on Performance

4 Research Design for Testing the BSC's Performance Improvement Theory

In order to empirically and statistically test the balanced scorecard's performance improvement theory, real world data is necessary. Since it is difficult to conduct experiments with actual organisations, most empirical studies on the balanced scorecard use either data derived from the organisation's information system or gathered through interviews. PWC (2001), for example, used questionnaires, which were sent to the top 200 German companies. Malina and Selto (2001) interviewed individuals within one company to obtain the data. Denton and White (2000) referred to the company's records when they judged the balanced scorecard project a success.

Common problems of these data sources include relevance of the information to the present, restricted availability due to the obligation to maintain confidentiality, and various dangers of biases when using interview-based approaches (Bortz and Döring 2002, 326-329). However, even if all desirable data for a broad range of organisations including balanced scorecard users and non-users were available, one problem would remain: the complexity of today's business organisations acting within a highly complex social economic system. It seems to be impossible to control or at least gather

information about all the independent variables which might influence an organisation's performance. For example, if a company introduces the balanced scorecard, activity based costing and six sigma quality philosophy at the same time, it will be difficult to relate changes in performance to the true cause. Moreover, one can not rule out the possibility that the change in performance is not at all caused by one of the measures taken but triggered by changes in the environment.

Although Bortz and Döring (2002, 61) rank experimental field studies first in internal and external validity, the above discussed issues prevent the application of this research method for testing the balanced scorecard's theory. In contrast to most other published studies, which use a quasi-experimental field design, I use an experimental laboratory approach as research design. The laboratory allows to gain control over almost all disruptive factors that might influence the dependent variable and results therefore in a very high internal validity (Bortz and Döring 2002, 60). External validity of the laboratory design might be questionable, as the artificiality of the laboratory might prevent the results from being honestly generalized. However, external validity could be improved by designing the experiment carefully and as realistically as possible.

Early in the design process of the experimental investigation it became obvious that testing all 13 hypotheses listed in section 3 with only one experiment would be too ambitious. As core hypothesis, H_1 could not be omitted. Because of Kaplan and Norton's emphasis on the strategic relevance of the balanced scorecard, H_{11} was chosen as a second hypothesis to be tested. Having narrowed the focus, the task was to design an experiment that would provide data which would allow to answer the question whether a balanced scorecard improved the implementation of an organisation's strategy and – as a result – its performance.

In the actual business environment the users of a balanced scorecard are the members of an organisation. The individuals with "strategic" power and able to implement a strategy in an organisation are the top managers. They receive and process the information provided by a scorecard; based on this information, they decide on measures and give orders. To be realistic, participants in the experiment therefore have to act as top managers. They are given the role of a managing director of a recently founded restaurant business – Happy Family Restaurants (HFR). HFR's business concept, strategy and environment are described in detail in a 13 page case-study. The participant's first task is to carefully read the essay and absorb as much information as possible. He is to learn that HFR's strategic goal is to grow sales revenue from 1 Million € to 300 Million € within a ten year time-frame, while maintaining profitability throughout the period. Return on capital (ROC) is to be greater or equal to 15 % per annum. The essay includes the HFR strategy paper that discusses 14 strategic issues, which are regarded as important for successfully implementing the growth strategy. It also shows causal links between the 14 strategic issues, providing the reader with something close to a strategy map. Each strategic issue is operationalised by one to three measures for which the long-term goal and the actual value is given. Finally, the case study describes the variables which the participants can change when implementing the growth strategy.

The HFR case-study is handed out to the participants one week before the experiment. It is the only input given in advance. The experiment is conducted as a computer aided simulation experiment following similar research conducted by Dörner et al (1994),

Akcermann et al (1995), Wittmann et al (1995) or Größler (2000). A simulator specifically developed for the HFR case is used. The core of the simulator is a System Dynamics simulation model built with Vensim. The user interface is programmed in Delphi. The simulator is similar to the fairly well-known Beefeater Restaurants Microworld by Global Strategy Dynamics Ltd.

Depending on the experiment setting, participants decided on nine or four parameters while implementing HFR's strategy (Figure 6). In the high-complexity setting the simulator's user interface allowed participants to change nine parameters on a quarterly basis. These parameters included more operational ones, such as target price of the average meal, and more strategic ones, such as expenditure for new restaurant sites. The low-complexity setting reduced the number of parameters, which the user could influence, to the four more strategic ones. The operational parameters were fixed on their optimal value.

All nine parameters influenced variables that related to the 14 strategic issues and finally had an effect on the performance measures ROC and sales revenue. Having made their decisions, participants could continue by simulating one quarter ahead. The outcomes of their decisions were computed, and the cumulated deviation between actual and desired values of the performance measures was displayed in the window "Zielerreichung". Since HFR's strategy horizon was ten years, participants could play a maximum of 40 quarters. They had the task to do their best in implementing HFR's strategy: to maximize the positive deviation from their strategic goals.

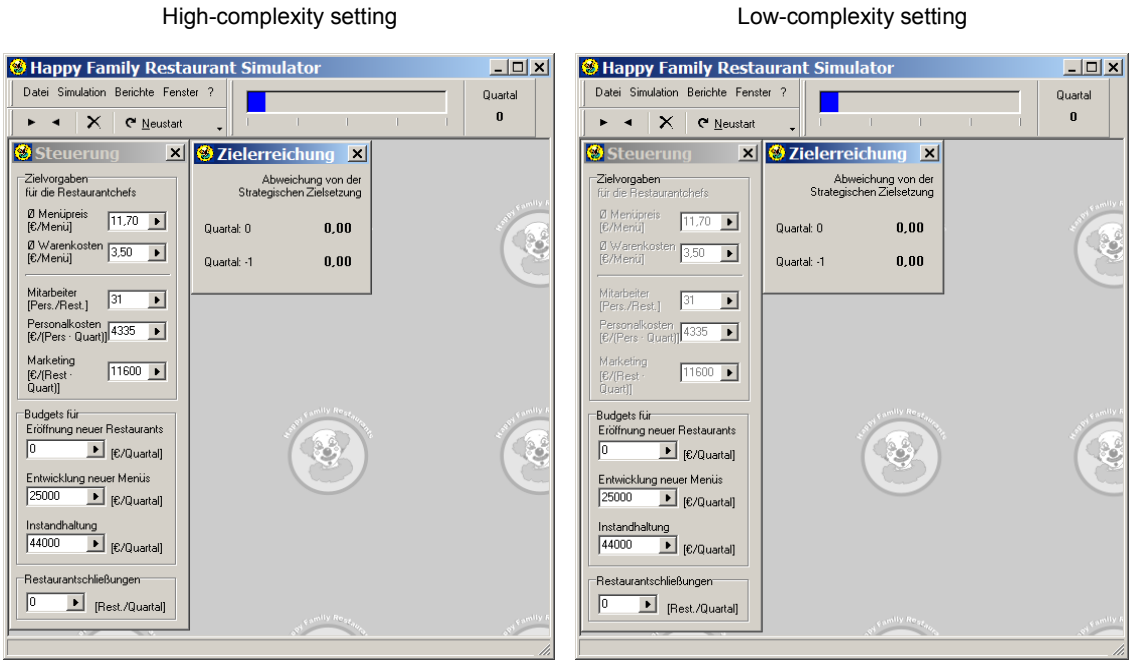

Figure 6: Screenshot of the Happy Family Restaurant Simulator

The HFR simulator allows to report the simulation results to the individual participant by two different means: traditional reports, as shown in Figure 7, and a balanced scorecard, as shown in Figure 8. A post-test only randomised experimental design with two groups is used as research design (Trochim 2002)). In the experimental laboratory the participants are divided by chance into two groups: the program group has only access to the balanced scorecard report and does not have the traditional form available;

the control group is equipped with simulators that only show the traditional reports. Thus, this research setting operationalized H_1 as follows:

H_{1O}: Participants in the laboratory experiment using a BSC as management and controlling system will perform better than participants using traditional reports as management and controlling system.

This operationalization permitted investigation of only one specific means how the BSC can influence organizational performance. Other benefits of the BSC stated by Kaplan and Norton (1996a, p.73), such as performance improvement through better strategy alignment, were deliberately excluded by the chosen research design.

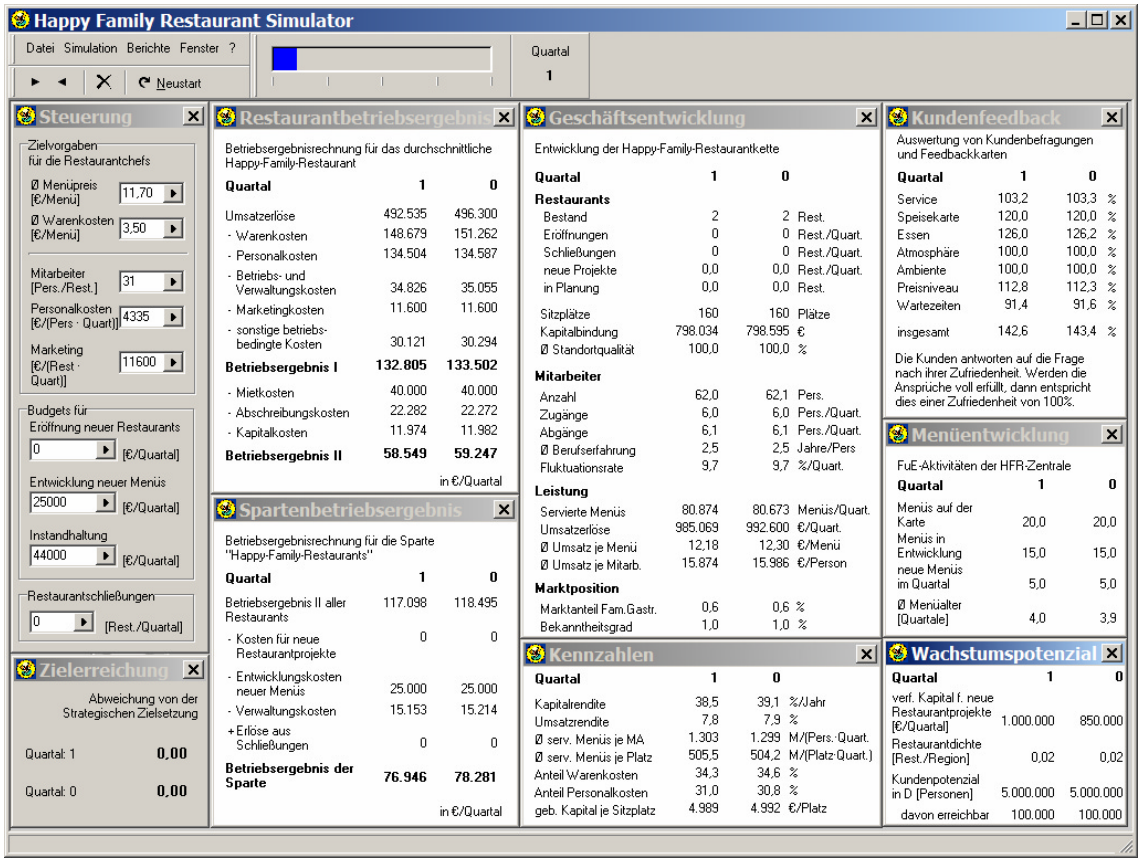

Figure 7: HFR Simulator's Traditional Reports

Participants in both groups are given the same time frame: 90 minutes for doing their best in implementing HFR's strategy, that is to maximize the positive cumulated deviation from their strategic goals. If they fail, which means having a cumulated deviation of –50 or less, they are laid off. However, participants are allowed to restart the simulation as often as they wish. The number of simulation runs and the duration of each simulation are recorded together with all other results in the simulation data file. Upon completion of the time span, the data files with the simulation results are collected so that the relevant data can be extracted.

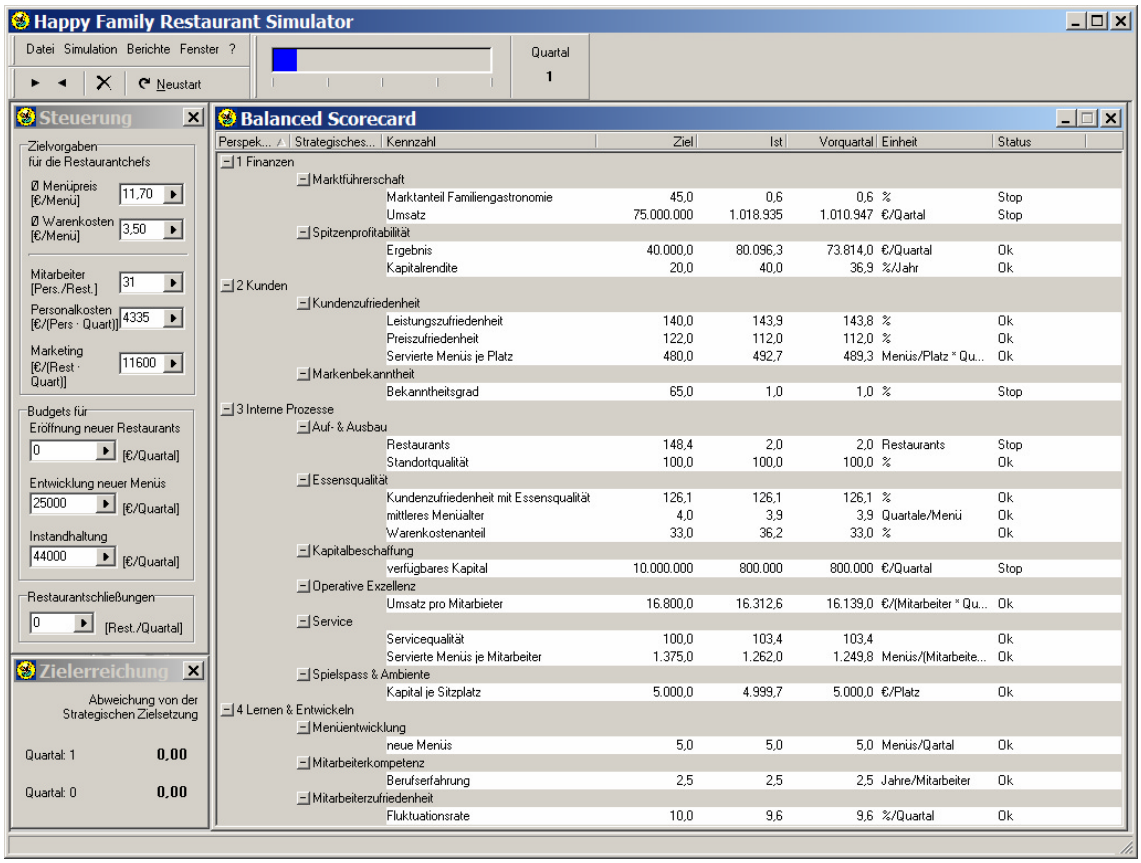

Figure 8: HFR Simulator with Balanced Scorecard Report

Because of the true experimental design used, results should be of high internal validity (Trochim 2002). Nevertheless the research design's sophistication was further enhanced to get additional potential influencing factors under control. The German psychologists Wittmann et al. (1995) and Süß (1996), for example, find strong relationships between a participant's intelligence, his general knowledge of business and economics, his computer game related system knowledge and the overall performance in a computer game. Figure 9 shows an excerpt from a later published path model with correlations.

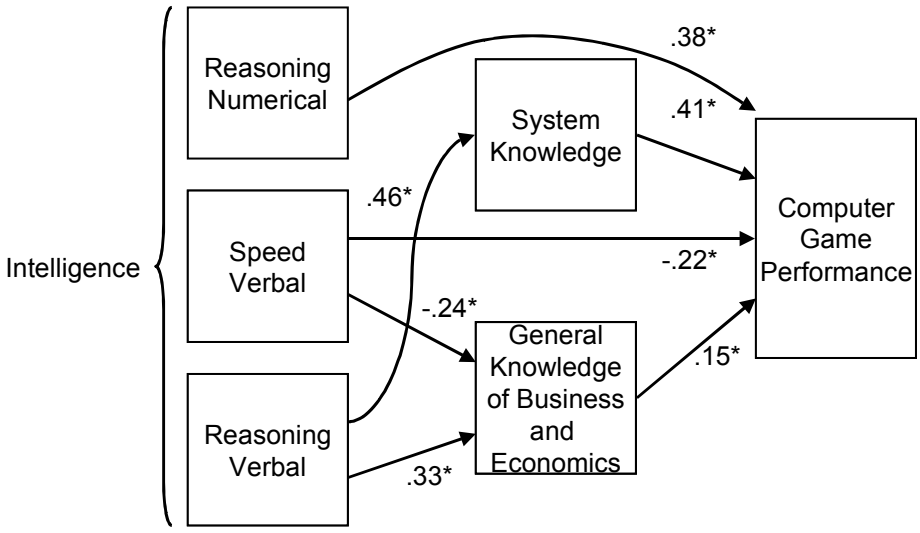

Figure 9: Nomological Network for Explanation of Complex Problem Solving – The Berlin 1989 Study (Wittmann/Hattrup 2004)

Ackerman and Kanfer (1993), Ackerman et al. (1995) and Ackerman and Cianciolo (2002) have undertaken similar research with similar results. They repeatedly found significant relationships between intelligence and performance in a Terminal Radar Air Control simulation. These results are consistent with the earlier meta-analytic work by Hunter and Hunter (1984), which showed that intelligence is a stronger predictor of overall job performance as the cognitive demands of job tasks increase.

The findings discussed above are used as an opportunity to administer two tests and one questionnaire before the simulation experiment. This allows to test whether factors other than the information system have a significant performance impact. First, a questionnaire is used to gather data about the participants' gender, grades in business administration courses taught at the HfB and to obtain subjective assessments of their knowledge about the balanced scorecard and the HFR business. Second, a knowledge test is performed, which consists of 32 multiple choice questions measuring the participant's general knowledge of business and economics. Third, the HFR specific knowledge of the participants is tested. Therefore, a short multiple choice test with 10 questions about HFR's situation and causal relationships in the restaurant business is used. A thorough reading of the case study should enable the subjects to answer all 10 questions correctly. Forth, the intelligence of the participants is tested with carefully selected exercises from the BIS-4 test (Jäger et al. 1997). The Berlin Model of Intelligence (BIS) disaggregates human intelligence using two dimensions: operations and content. The intelligence dimension "operations" includes mental speed (B), short term memory (M), creativity (C) and reasoning (K); the intelligence dimension "content" includes figural/spatial (F), verbal (V) and numerical (N). Since Wittman et al do not find significant correlations between M, C and computer game performance, these two dimensions are omitted. In total, 15 exercises are used to derive measures for B and K as well as F, V, and N. The time required to deal with the exercises adds up to 45 minutes.

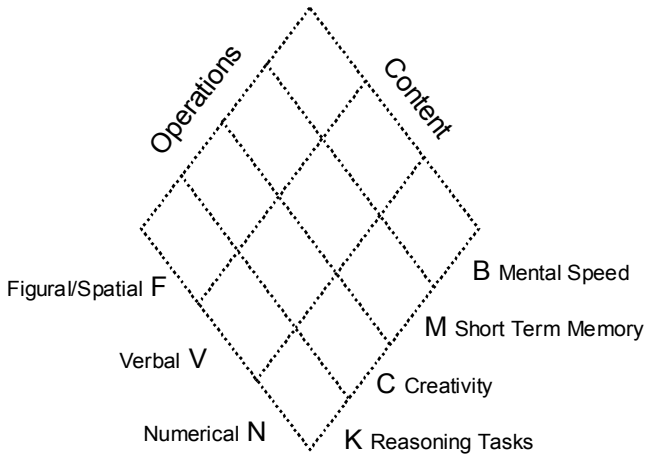

Figure 10: The Berlin Model of Intelligence (Jäger et al 1997, 5)

As a result, the total time required for one laboratory experiment is 195 minutes. After a short debriefing, it starts with answering the general questionnaire (5 min). The two knowledge multiple choice tests take 20 minutes and 5 minutes respectively. After the intelligence test (45 min), a 15-minutes break is taken. Finally, the simulation experiment itself requires 90 minutes.

As shown in Figure 11, the data gathered by three tests and the questionnaire allows for enhanced analysis of a dozen additional potential performance drivers. This clearly improves the internal validity of the research design chosen.

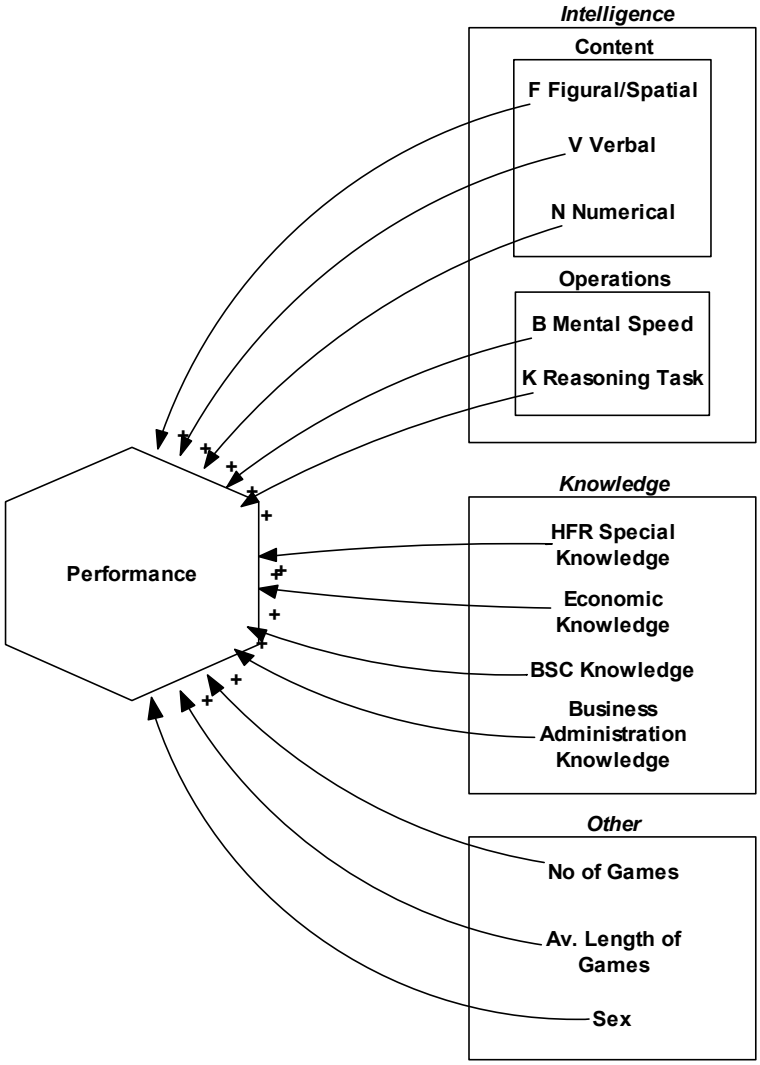

Figure 11: Other Performance Drivers

5 Implementation of the Research Design

Between May 2003 and January 2004 a total of eight experiments were performed, each involving 18 to 33 participants. The experiments were conducted with students in their final semester in the Bachelor of Business Administration and Diploma programs at the HfB – Business School of Finance and Management. Since strategy implementation using the balanced scorecard is performed by top managers, who have usually attended university, Bortz and Döring's (2002) reservations about students as subjects are not accurate for the purpose of this study. For the same reason is can be considered unproblematic that the average intelligence of the participants (mean: 105.72, standard deviation: 11.24) lies above the BIS test's reference (mean: 100, standard deviation: 10).

As the true experimental design requires, the participants were assigned to the experiment group and the control group by chance. This was done by drawing

identification numbers from a pool of 50 numbers for each experiment. 50 % of these numbers were corresponded to a HFR simulator equipped with a balanced scorecard (treatment value = 2); the other 50 % were related to a simulator interface showing traditional reports (treatment value = 3). As Figure 12 illustrates, except for experiments five, six and seven, participants were allocated almost equally to the experiment and control groups.

Experiment	Date	Course	Treatment	No. Participants
1	2003-05-22	8A1	2 (BSC)	14
			3 (Reports)	12
2	2003-05-27	8B2	2 (BSC)	14
			3 (Reports)	11
3	2003-06-03	8B1	2 (BSC)	12
			3 (Reports)	13
4	2003-11-27	7DO2	2 (BSC)	18
			3 (Reports)	15
5	2003-12-04	7DO1	2 (BSC)	21
			3 (Reports)	11
6	2003-12-08	7MO1	2 (BSC)	21
			3 (Reports)	8
7	2003-12-12	9AIS	2 (BSC)	11
			3 (Reports)	7
8	2004-01-06	7BL1	2 (BSC)	12
			3 (Reports)	14

Figure 12: Number and Allocation of Participants

Because participants were randomly assigned to two groups in all eight experiments, it can be assumed that the two groups are probabilistically equivalent (Trochim 2000). Nevertheless, using the results from the questionnaire, the knowledge tests and the intelligence tests, this assumption can be tested.

In the questionnaire, participants were asked to give their best estimate concerning their average exam results in Business Administration courses at the HfB as a percentage of the maximum value (BAK). The knowledge value about the concept of the balanced scorecard was derived from four equally weighted questions. On the one hand, participants were asked to rate their overall balanced scorecard knowledge (BSC1) using a scale from 1 to 5; on the other hand, they were asked how carefully they had read each of the three articles about the balanced scorecard (BSC2-BSC4). This was done using a scale from 1 to 6. The knowledge measure about the balanced scorecard (BSCK) was finally computed with the following equation:

$$BSCK = \frac{BSC1 \cdot 2 + BSC2\,\frac{5}{6} + BSC3\,\frac{5}{6} + BSC4\,\frac{5}{6}}{5}$$

The general economic and business knowledge (GEK) of the participants was measured using the percentage of correct answers in the knowledge multiple choice test. The participants' specific HFR knowledge (HFRK) was not determined for experiments one, two and three. Beginning with experiment four a second multiple choice test was used. Like the general economic and business knowledge, the specific knowledge was measured with the percentage of correct answers. Figure 13 summarizes mean (μ) and standard deviation (σ) for the four measures described above.

Experiment	Treatment	BAK		BSCK		GEK		HFRK	
		μ	σ	μ	σ	μ	σ	μ	σ
1	2	75.00%	10.13%	3.92	0.75	72.66%	6.68%		
	3	72.50%	9.03%	4.15	0.78	73.58%	7.95%		
2	2	73.67%	7.53%	4.00	0.68	71.35%	6.64%		
	3	75.42%	6.65%	3.85	1.00	74.69%	8.65%		
3	2	72.22%	9.05%	3.78	0.84	68.49%	5.41%		
	3	70.45%	14.30%	3.71	1.12	72.84%	9.23%		
4	2	76.80%	5.97%	4.05	0.86	71.70%	8.27%	63.89%	11.95%
	3	73.00%	9.41%	4.01	0.59	67.71%	9.12%	62.67%	12.80%
5	2	74.65%	9.21%	4.27	0.66	66.82%	19.17%	61.90%	14.70%
	3	72.56%	10.19%	4.50	0.36	71.02%	13.12%	63.33%	10.00%
6	2	72.31%	7.54%	3.49	0.91	65.16%	6.51%	64.21%	9.61%
	3	77.86%	9.94%	4.13	0.91	73.05%	10.95%	50.00%	16.90%
7	2	73.03%	11.87%	3.95	0.98	72.73%	8.96%	53.64%	15.02%
	3	63.57%	9.20%	4.43	0.31	68.30%	12.81%	41.67%	14.72%
8	2	70.60%	11.64%	4.01	0.54	70.31%	8.89%	60.83%	16.76%
	3	76.14%	10.84%	3.89	0.76	68.53%	7.99%	64.29%	10.16%

Figure 13: Basic Descriptive Statistics of the Participants' Knowledge

The intelligence of the subjects was measured with the BIS intelligence test. Figure 14 shows mean and standard deviation for the standardized overall intelligence (AI_S), grouped by experiment and treatment.

Experiment	Treatment	AI_S		Experiment	Treatment	AI_S	
		μ	σ			μ	σ
1	2	107.5	15.0	5	2	100.8	12.3
	3	103.9	11.4		3	109.1	9.4
2	2	102.8	9.6	6	2	103.5	9.9
	3	109.6	12.7		3	103.8	13.7
3	2	107.7	10.0	7	2	105.7	9.6
	3	102.4	9.4		3	112.0	15.5
4	2	108.1	9.1	8	2	108.7	12.2
	3	105.0	11.8		3	107.7	9.6

Figure 14: Descriptive Statistics of the Subjects' Intelligence

In order to obtain valid experimental results, treatment group (2=balanced scorecard) and the control group (3=reports) should not differ systematically regarding knowledge and intelligence. If significant differences existed between the groups, discrepancies in the simulation performance could not be attributed to the information system available. Therefore the groups were tested for significant differences. The Independent-Samples T Test is commonly used for this purpose. It requires, however, that the analysed variables are normally distributed. Therefore the Kolmogorov-Smirnov Test, which compares the observed cumulative distribution function for a variable with the normal distribution, was executed first.[2] Figure 16 shows the outcome.

[2] The SPSS software package is used for statistical testing.

	Treatment	BAK	BSCK	GEK	HFRK	AI_S
N	2	100	115	118	81	118
Kolmogorov-Smirnov Z		1.0587	2.1628	1.4175	1.3687	0.7589
Asymp. Sig. (2-tailed)		0.2123	**0.0002**	**0.0360**	**0.0472**	0.6122
N	3	83	80	89	52	88
Kolmogorov-Smirnov Z		0.8817	1.5190	0.9062	1.3380	0.5304
Asymp. Sig. (2-tailed)		0.4185	**0.0198**	0.3842	0.0557	0.9411

Figure 15: Kolmogorov-Smirnov Test (Aggregated Data)

Figure 15 indicates that BSCK, GEK and HFRK can not be assumed normally distributed when aggregating the data to the treatment group level. When, however, the data of each experiment and treatment group are tested separately, no significant difference from the normal distribution occurs (Figure 16).

Experiment	Treatment	BAK	BSCK	GEK	HFRK	AI_S
1	2	0.988	0.772	0.977		0.987
	3	0.490	0.322	0.856		0.619
2	2	0.934	0.991	0.943		0.970
	3	0.976	0.797	0.927		0.872
3	2	0.760	0.821	0.712		0.820
	3	0.869	0.390	0.179		0.743
4	2	0.774	0.379	0.875	0.257	0.798
	3	0.302	0.967	0.647	0.477	0.589
5	2	0.376	0.374	0.259	0.574	0.897
	3	0.570	0.681	0.998	0.895	0.919
6	2	0.658	0.302	0.690	0.192	0.676
	3	0.556	0.679	0.919	0.993	0.799
7	2	0.500	0.082	0.863	0.201	0.825
	3	0.869	0.996	0.974	0.946	0.919
8	2	0.967	0.608	0.941	0.261	0.870
	3	0.938	0.346	0.610	0.549	0.952

Figure 16: Asymp. Sig. (2-tailed) of the Kolmogorov-Smirnov Test

Because the Kolmogorov-Smirnov Test produced insignificant results, the T Test could be executed. As Figure 17 illustrates, the significance values reach a critical limit for very few variables and experiments, which means that the equality of variances and means should be rejected. To complete the analysis, the Mann-Whitney U test was performed with the purpose to test whether two sampled populations are equivalent in location. The only data requirement is that the two samples are similar in shape. Normal distribution is not presupposed. Figure 18 shows the results. The Mann-Whitney significance level is below 0.05 for five experiment-variable-pairs, which means that the null hypothesis of equal groups has to be rejected. When aggregating the data of all experiments, Mann-Whitney does not indicate significant differences between the groups.

Experi-ment	Significance	BAK	BSCK	GEK	HFRK	AI_S
1	Levene's Test Sig.	0.626	0.681	0.567		0.167
	T-TestSig. (2-tailed)	0.530	0.461	0.765		0.505
2	Levene's Test Sig.	0.926	0.601	0.489		0.357
	T-TestSig. (2-tailed)	0.613	0.173	0.318		0.137
3	Levene's Test Sig.	0.061	**0.005**	0.505		0.712
	T-TestSig. (2-tailed)	0.752	0.212	0.169		0.186
4	Levene's Test Sig.	0.365	0.871	0.899	0.800	0.578
	T-TestSig. (2-tailed)	0.197	0.380	0.197	0.779	0.408
5	Levene's Test Sig.	0.840	0.265	0.449	0.220	0.556
	T-TestSig. (2-tailed)	0.600	0.644	0.521	0.793	0.076
6	Levene's Test Sig.	0.598	**0.022**	0.069	0.137	0.424
	T-TestSig. (2-tailed)	0.155	0.455	**0.025**	**0.010**	0.962
7	Levene's Test Sig.	0.124	**0.015**	0.579	0.845	0.059
	T-TestSig. (2-tailed)	0.093	**0.028**	0.399	0.135	0.300
8	Levene's Test Sig.	0.779	0.611	0.492	**0.046**	0.486
	T-TestSig. (2-tailed)	0.244	0.061	0.594	0.542	0.818

Figure 17: Significance Values of Leven's Test for Equality of Variances and T Test for Equality of Means

Experiment	BAK	BSCK	GEK	HFRK	AI_S
1	0.392	0.440	0.802		0.487
2	0.527	0.176	0.351		0.139
3	0.698	0.359	**0.029**		0.264
4	0.290	0.315	0.265	0.985	0.492
5	0.427	0.856	0.796	0.745	0.108
6	0.147	0.634	**0.040**	**0.033**	0.690
7	0.169	**0.006**	0.522	0.183	0.413
8	0.346	**0.042**	0.567	0.831	0.579
Total (1-8)	0.562	0.846	0.277	0.454	0.545

Figure 18: Mann-Whitney Asymp. Sig. (2-tailed)

All in all, the test results show that the true experimental design resulted in fairly equal treatment and control groups. Some significant inequalities, however, indicate that the additional tests and questionnaires might not be useless. The enhanced sophistication of the research design allows in any case a more complete analysis of the performance impact of the balanced scorecard.

6 A First Test of Balanced Scorecard's Performance Impact

Although the experiments have not yet been fully analysed and evaluated, first results are presented. The subjects' performance in the simulation experiment was measured with a single figure, which computes the aggregate deviation from the two strategic goals: ROC and sales revenue. Since the subjects had the possibility to perform multiple simulations, only the best run was chosen. Cancelled simulations and runs that ended with a layoff were valued –50. Figure 19 shows some descriptive statistics for the eight experiments.

Experiment	Treatment	N	Mean	Std. Deviation	Kolmogorov-Smirnov Asymp. Sig. (2-tailed)
1	2	14	-49.998	0.006	**0.0007**
	3	12	-42.922	17.370	**0.0061**
	Total	26	-46.733	12.071	
2	2	14	-32.787	22.022	0.5236
	3	11	-31.198	20.372	0.6336
	Total	25	-32.088	20.887	
3	2	12	-33.772	22.865	0.1161
	3	13	-37.189	18.308	**0.0533**
	Total	25	-35.549	20.255	
4	2	18	-42.783	20.663	**0.0101**
	3	15	-44.144	11.528	0.1211
	Total	33	-43.402	16.895	
5	2	21	-40.870	14.418	**0.0022**
	3	11	-35.617	16.735	0.1342
	Total	32	-39.064	15.195	
6	2	21	-44.066	12.732	**0.0001**
	3	8	-44.841	10.386	0.0899
	Total	29	-44.279	11.953	
7	2	11	-42.286	14.464	**0.0340**
	3	7	-27.629	21.441	0.6418
	Total	18	-36.586	18.422	
8	2	12	-33.762	20.117	0.0702
	3	14	-27.676	20.671	0.1943
	Total	26	-30.485	20.244	

Figure 19: Simulation Performance Descriptive Statistics

It can be noticed that the mean performance of the treatment group with a balanced scorecard available is higher than in the control group only in experiments three, four and six. In all other experiments, the subjects using traditional reports show a higher performance. Nevertheless, it is unclear whether these differences are statistically significant.

To test the normal distribution prerequisite of the t test, the Kolmogorov-Smirnov test was used again. The results – shown in Figure 19 as well – indicate that for half of the experiment groups no normal distribution of the performance measure can be assumed. Since an important assumption of the t test was not met, two nonparametric tests for two independent samples were used to determine whether or not the performance differed significantly between the two groups. Mann-Whitney's and Kolomogorv-Smirnov's tests show both that the null hypothesis – both groups are equal in their simulation performance – can not be rejected. The significant values are much too high. As a first result, one can state that the balanced scorecard neither increased nor decreased the performance significantly when compared with traditional reports.

However, before definitely rejecting hypotheses H_1 and H_{11}, a more sophisticated and advanced statistical analysis has to be performed.

Experiment	Mann-Whitney Asymp. Sig. (2-tailed)	Two-Sample Kolmogorov-Smirnov Asymp. Sig. (2-tailed)
1	0.4042	0.9939
2	0.7773	0.9947
3	0.7588	0.9921
4	1.0000	1.0000
5	0.3736	0.8578
6	0.8904	0.9998
7	0.1353	0.5347
8	0.4416	0.9280

Figure 20: Mann-Whitney and Two-Sample Kolmogorov-Smirnov Significance Levels Regarding the Simulation Performance

7 Conclusions

Testing the balanced scorecard's built-in theory about its impact on organisational performance is of high theoretical and practical interest. It is possible to formulate a system of hypotheses which explains clearly the various effects of the balanced scorecard on performance. Testing this hypotheses system with a true field experiment is, however, for various reasons impossible. Therefore a true laboratory experiment was designed. With a realistic case study, a computer simulated micro-world and a carefully designed research process the experiment could be given a high external validity. Internal validity– high anyway due to the random assignment of subjects to the two groups – was improved by gathering data about a broad range of other possible influence factors on performance. At the same time this opens up further possibilities of enhanced causal analysis.

Preliminary statistical analysis indicates that the balanced scorecard's impact on performance might be overestimated. However, before a reliable final statement can be given, enhanced statistical analysis has to be performed. Possible limitations of the research have to be analysed as well. Süß (1996), for example, explains supposedly insignificant relationships between intelligence and performance by inappropriate performance measures and overtaxed participants. Further research has to rule out that these effects have indeed occurred.

8 References

Ackerman, P. L./Cianciolo, A. T.: Ability and Task Constraint Determinants of Complex Task Performance, in: Journal of Experimental Psychology: Applied, 8(3), 2002, pp. 194-208.

Ackerman, P. L./Kanfer, R./Goff, M.: Cognitive and Non-Cognitive Determinants and Consequences of Complex Skill Acquisition. Journal of Experimental Psychology: Applied, 1, 1995, p. 270-304.

Ackerman, P. L./Kanfer, R.: Integrating Laboratory and Field Study for Improving Selection: Development of a Battery for Predicting Air Traffic Controller Success. Journal of Applied Psychology, 78, 1993, p. 413-432.

Argyris, Chris: On Organizational Learning, 2nd Ed., Oxford: Blackwell, 1999.

Argyris, Chris: Single-Loop and Double-Loop Models in Research on Decision-Making, in: Administrative Science Quarterly, Vol. 21, Iss. 3, September 1976; p. 363.

Atkinson, Anthony et al.: New Directions in Management Accounting Research, in: Journal of Management Accounting Research, Vol. 9, Iss. 1, 1997, p. 79-108.

Denton, Gregory A./White, Bruce: Implementing a Balanced-Scorecard Approach to Managing Hotel Operations, in: Cornell Hotel & Restaurant Administration Quarterly, 2000, Vol. 41 Issue 1, pp. 94-108

Dörner, Dietrich et al. (Hrsg.): Lohhausen: Vom Umgang mit Unbestimmtheit und Komplexität, unveränderter Nachdruck, Bern u.a.: Huber, 1994.

Forrester, Jay W.: Industrial Dynamics, Cambridge, Mass.: The M.I.T Press, 1961.

Goulian, Caroline/Mersereau, Alexander: Performance Measurement: Implementing a Corporate Scorecard, Ivey Business Journal, Sep/Oct 2000; Vol. 65, Iss. 1; p. 48-54

Größler, Andreas: Entwicklungsprozess und Evaluation von Unternehmenssimulatoren für lernende Unternehmen, Frankfurt am Main et al.: Peter Lang, 2000

Jäger, Adolf Otto/Süß, Heinz-Martin/Beauducel, André: Berliner Intelligenzstruktur-Test: BIS-Test Form 4: Handanweisung, Göttingen et al: Hogrefe, 1997

Kaplan, Robert S./Norton, David P.: Balanced Scorecard: Translating Strategy into Action, Boston, Mass.: Harvard Business School Press, 1996b

Kaplan, Robert S./Norton, David P.: How Strategy Maps Frame an Organization's Objectives, in: Financial Executive, Vol. 20 Issue 2, pp. 40-46, Mar/Apr 2004a.

Kaplan, Robert S./Norton, David P.: Strategy Maps: Converting Intangible Assets Into Tangible Outcomes, Boston, Mass.: Harvard Business School Press, 2004b.

Kaplan, Robert S./Norton, David P.: Strategy-Focused Organization: How Balanced Scorecard Companies Thrive in the New Business Environment, Boston, Mass.: Harvard Business School Press, 2000

Kaplan, Robert S./Norton, David P.: The Balanced Scorecard – Measures that Drive Performance, in: Harvard Business Review, January-February 1992, p. 71-79.

Kaplan, Robert S./Norton, David P.: Transforming the Balanced Scorecard from Performance Measurement to Strategic Management: Part I, in: Accounting Horizons, Vol. 15, No. 1, March 2001, pp. 87-104.

Kaplan, Robert S./Norton, David P.: Using the Balanced Scorecard as a Strategic Management System, in: Harvard Business Review, January-February 1996a, pp. 75-85.

Kim, Daniel H.: The Link Between Individual and Organizational Learning, Sloan Management Review, Vol. 35, Issue 1, Fall 1993, p.37-51

Klein, Barbara D.: User Perceptions of Data Quality: Internet and Traditional Text Sources, in: Journal of Computer Information Systems, Vol. 41, Issue 4, Summer 2001, p 9-16.

Malina, Mary A./Selto, Frank H.: Communicating and Controlling Strategy: An Empirical Study of the Effectiveness of the Balanced Scorecard, in: Journal of Management Accounting Research, 2001, Vol. 13, pp. 47-91

Miller, George A.: The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information, in: The Psychological Review, Vol. 63, 1956, p. 81–97.

Mooraj, Stella/Oyon, Daniel/Hostettler, Didier: The Balanced Scorecard: A Necessary Good or an Unnecessary Evil?, in: European Management Journal, Vol. 17, Iss. 9, Oct, 1999, p. 481-489.

Nørreklit, Hanne: The Balance on the Balanced Scorecard - A Critical Analysis of Some of its Assumptions, in: Management Accounting Research, Vol. 11, Iss. 1; March 2000, p. 65-88.

PWC Deutsche Revision (Ed.): Die Balanced Scorecard im Praxistest: Wie zufrieden sind die Anwender?, Frankfurt am Main, 2001, (http://www.pwc.de/30000_publikationen/meldung.asp?id=224, 29.05.02)

Rigby, Darrell: Management Tools and Techniques: A Survey, in: California Management Review, Vol. 43, No. 2, Winter 2001, p. 139-160.

Sterman, John D.: Business Dynamics: Systems Thinking and Modeling for a Complex World, Boston et al: Irwin McGraw-Hill, 2000.

Strohhecker, Jürgen: Leitmotive des strategischen Bankmanagements, in: Udo Steffens/Wieland Achenbach: Strategisches Management in Banken, Frankfurt: Bankakademie-Verlag, 2002, S. 7-29.

Trochim, William: The Research Methods Knowledge Base, 2nd Edition, Cincinnati, OH: Atomic Dog Publishing, 2000.

Wang, Richard Y. /Strong, Diane M.: Beyond Accuracy: What Data Quality Means to Data Consumers, in: Journal of Management Information Systems, Vol. 12, Iss. 4, Spring 1996, pp. 5-33;

Witmann, Werner et al.: Der Zusammenhang von Arbeitsgedächtniskapazität und Konstrukten der Intelligenzstrukturforschung, in: Berichte des Lehrstuhls Psychologie II, Heft 1, 1995

Wittmann, Werner W./Hattrup, Keith: The Relationship Between Performance in Dynamic Systems and Intelligence, in: Systems Research and Behavioral Science, Vol. 21, 2004, 393-409.