Design and validation of a quality scan for the primary education
the relation between school related factors and student performance

van Rooij, N.

Award date:
2009

Eindhoven, January 2009

Design and Validation of a quality scan for the Primary Education: The relation between school related factors and student performance

By
N. (Nick) van Rooij

Student identity number 0551351

in partial fulfilment of the requirements for the degree of

Master of Science
in Operations Management and Logistics

Supervisors
TU/e Dr. Ir. P.A.M. Kleingeld, TU/e, HPM
Dr. A. De Jong, TU/e, OSM
Company Dr. M. Peeters, OCGH Advies

TUE. Department Technology Management.
Series Master Theses Operations Management and Logistics

ARW 2009 OML

Subject headings: Education, Survey, Factor Analysis, Statistical Data Analysis, Teacher, School Performance, Basis-onderwijs, Onderwijsonderzoek, Onderwijsevaluatie, Enquête

Preface

This report describes the result of my graduation project of the Master Operations Management and Logistics at the Eindhoven University of Technology. With this project, I have finished a period consisting of five and a half year of study at Eindhoven. The project has been executed for the capacity group Human Performance Management and was conducted at OCGH Advies at Helmond.

I would like to thank Ad Kleingeld for his advice and effort during the project. The appointments almost always took more time than planned, but this did not concern him. In addition, I want to thank Ad de Jong for his insights in the analysis. Next, I like to thank Anniek van Bemmelen for her help with the translation of the questionnaire into Dutch.

Furthermore, I express my thanks to all employees of OCGH Advies for the pleasant cooperation during my project. I want to name my supervisor Miranda Peeters especially since each week she made some time for me to discuss the progress of my study and where necessary provided me with advice. Of course I also thank the director Ad de Veer for letting me complete the project at OCGH Advies.

Subsequently, I would like to thank Rick Derks. He checked almost all texts for both my literature study as this report on English. In addition, I like to thank Rutger Stultiëns, for checking my English in a few chapters of this report.

Finally, I thank my parents for their support during my study.

Nick van Rooij,
Helmond, January 2009.

Content

Summary

This report represents the results of the graduation project for the Master Operations Management and Logistics (OML) at Eindhoven University of Technology. The project is conducted at the education consultancy agency OCGH Advies at Helmond.

Introduction

The project focused on the Primary Education in the Netherlands. Nowadays, the Dutch Government provides schools with more freedom concerning the organization of their education activities. Nevertheless, the Government performs a monitoring role in the Primary Education, which is executed by the Dutch Education Authority. It monitors the quality at the Dutch schools using a standardized evaluation tool on twelve different factors. It is therefore valuable to a school to know their quality at several educational aspects.

OCGH Advies is an education consultancy agency with about 50 employees, which is mainly active in the district of Zuidoost Brabant. OCGH Advies offers advice and support to Primary and Secondary schools. This can be for the teachers, but also for the principal, the students, and their parents. In January 2008, it moved from being a subsidized organization to a company in the private sector. As a result, the organization now focuses more on the innovation of their range of products. This is because schools can now choose among several consultancy agencies for their supportive needs.

The project objective

The starting point for this project was the work of Marzano (2003). His book described eleven categories which could influence the student performance at school. The categories were divided into three levels: the school, teacher, and student level. Based on a meta-analysis, Marzano (2003) provided a questionnaire with different statements for each category. OCGH Advies wanted to examine whether the questionnaire would be applicable for the Primary Education in the Netherlands, since Marzano's results were mainly based on studies conducted in the United States. It therefore should be indicated whether the categories and statements were relevant in the Netherlands. Obviously, the statements should also be translated into Dutch. The project objective of the Master Thesis was formulated as follows:

Design and validate a quality scan instrument, which provides educational-related factors influencing student performance in the Primary Education.

Although the Master Thesis was mainly design focused, some empirical research questions were also formulated. These were divided into several groups. First, *generalisation* concerned whether adjustments should be made to the questionnaire of Marzano (2003) and what these adjustments should be. Second, *the properties of the quality scan* focused on the internal structure of the different statements of the quality scan. It examined the relations between the factors of the quality scan. Third, *relations with other instruments* investigated whether there was a relation between the quality scan and the EFQM model, and the evaluation tool of the Dutch Education Authority. The EFQM model is useful for all kinds of companies and it is a model generally used by Dutch schools to conduct a self-assessment. Finally, *prediction of student performance* focused on the relation between the quality scan and various measures of student performance.

Research design

The study was conducted in several phases, which made the project more orderly. The phases were not executed completely sequentially, but provided a good research model. The phases of the project were:

1. *Literature review*. Additional literature was studied to verify the accuracy of Marzano (2003) were correct or not. Besides, literature for two additional categories, leadership and teacher motivation, was studied.
2. *Translation of questionnaire.* The translation of the questionnaire into Dutch was conducted with the traditional translation and committee translation methods (Balbinotti et al., 2006).
3. *Information from OCGH Advies.* Information from OCGH Advies and information from the Primary Education in general were used as input for the reformulation of the quality scan.
4. *Design of the quality scan.* The translated questionnaire was analyzed and adjustments were made to the questionnaire.
5. *Data collecting.* The quality scan was completed by employees of several schools. These were used for the validation process of the quality scan.
6. *Analyzing data.* Different analyses were conducted to establish the structure of the quality scan and answer the research questions.
7. *Implementation.* A tool, based on Excel, was created to implement the data for each school. Besides, a feedback report was designed.

The research design can be divided in two main parts. First, the design of the quality scan was conducted during phases one to four. Second, the validation of the quality scan followed from phases five and six.

Design of the quality scan

The objective of the design of the quality scan was to develop a quality scan that could be used for the empirical study. This was done in five steps. Step one was the translation of the questionnaire into Dutch. Step two was the reformulation of the statements, so they would become more applicable in the Dutch Primary Education. Step three concerned the structure of the quality scan. During step four, a pilot test was conducted at one school (i.e. with 18 respondents), which provided useful information for the final adjustments which were made at step five. The results of these steps were as follows: the quality scan consists of eleven categories with 92 statements (with six statements for general information); In comparison with the questionnaire of Marzano (2003), 14 statements were deleted, 28 statements were added, and 6 statements were split up; In addition, four statements were reallocated to a different category; two categories were removed. As a result, 44 percent of the statements used in the quality scan had no direct relation with the original questionnaire of Marzano (2003). The changes were mainly due to the inclusion of the factors leadership and teacher motivation. Secondly, Marzano's questionnaire could also be used for the Secondary Education, and thus various statements were irrelevant for the Primary Education. Subsequently, changes were caused because of the differences in education in the United States and the Netherlands. The last reason for changes was that the statements were badly or ambiguously formulated.

Validation of the quality scan

The validation of the quality scan was conducted in a field study. The sample consisted of 137 participants from nine different schools. The independent variables were the statements of the quality scan. A goal of the validation process was to analyze if these statements represented several statistical

factors. In addition, four types of dependent variables were included, being the average Cito End scores of a school, the average level of Secondary Education the students followed after finishing their Primary Education, results of a national standard test for reading and mathematics, and the factors used by the Dutch Education Authority.

The validation was accomplished with different analyses. First, a reliability analysis was executed for assessing the consistency within the categories of the quality scan. All except one category had a Cronbach's Alpha above 0,60, which is the critical value for an explorative study (Hair et al., 2006). Second, an exploratory factor analysis (EFA) was carried out. The EFA was split in different EFA models, due to the relatively small sample size. The Pearson correlation matrix provided appropriate correlations. Additionally, intercorrelations were present, so the EFA could be executed, which resulted in nineteen factors. Although a few factors did not have an appropriate reliability all were used for the next analysis. This analysis was a confirmatory factor analysis (CFA). With this analysis, it could be indicated whether the model fit was appropriate. Like the EFA, the CFA was also split up, because of the small sample size. All CFA models had an adequate fit. Therefore, it was concluded that the nineteen factors, which are presented in Table S1, could be used. The factors are divided into the school level, teacher level, and an extra classification.

School level	**Teacher level**	**Extra**
Learning goals (school)	Teacher motivation	Student motivation
Team professionalism	Personal development	Leadership
Student care system	Pedagogical behaviour	
Education curriculum	Class differentiation	
School development goals	Class rules	
Parent involvement	Les planning	
Safe environment	Practicing content	
	Approach knowledge obtainment	
	Goals and feedback (class)	
	Homework	

Table S1: factors of the quality scan.

Subsequently, an aggregation analysis was conducted using the within-group interrater reliability (r_{wg}) and the interclass correlation coefficient (ICC(1)), since the dependent variables were at the school level. Based on the first method, it was concluded that all factors could be aggregated, but the results should be handled with care. Finally, using a Kendall's tau b correlation matrix, the relations with the dependent variables were established. Here, a first indication of the relation between the quality scan and the dependent variables was provided. However, no hard conclusions followed from this analysis, due to the small sample size. Furthermore, the factors of the Dutch Education Authority could not be used, because of the lack of variance between the schools in the sample.

Conclusions

In the discussion section, the validity of the quality scan has been described. Scandura and Williams' (2003) model consisting of the distinction between four types of validity was used to determine the validity of the quality scan. These types are statistical conclusion validity, construct validity, internal validity, and external validity. During the discussion of the construct validity, the differences between the results for factors based on the *group-referent* statements (e.g. in my school students receive feedback on their knowledge growth) and factors based on the *individual-referent* statements (e.g. I emphasise the importance of effort to students) have been discussed. These are based on the *referent-*

shift consensus model and *direct consensus model* respectively (Chan, 1998). It was argued that the statements using the referent-shift consensus model revealed larger differences between schools and therefore adequate ICC(1) values, due to the larger between-group variability. Moreover, several explanations were provided for the unexpected outcomes of some methods related to the construct validity. Based on several measures it was motivated that the external validity was not accomplished for the whole population. The schools in the sample have all been classified to a subpopulation, namely schools with students coming from a good social-cultural background (i.e. students with a "*leerlinggewicht*" 1,00 (Terugblik en resultaten, 2008)). Nevertheless, most schools in the district of OCGH Advies have been classified into the same subpopulation. The final conclusion regarding the validity of the quality scan was that the different facets of validity of the quality scan were demonstrated. Although not all criteria for each factor or statement have been met, it can be concluded that the factors of the quality scan are valid.

Subsequently, the answers of the research questions were provided. Regarding the *generalization*, it was concluded that the quality scan was applicable for the Dutch Primary Education. However, I need to remark that Marzano (2003) completed an adequate meta-analysis for establishing relevant educational factors, but his conversion of these factors to statements in his questionnaire was not executed adequately. With respect to the *properties of the quality scan*, the reliability of the educational factors was established. It followed that almost all factors were correlated with each other. The relation with the EFQM-model was also presented, which belongs to the third group of research questions, *relation with other instruments*. It followed that most of the factors were related to the *organizational oriented criteria* and less to *the result oriented criteria*. This was not a disadvantage, since the goal of the quality scan was to signify the educational processes at a school. The last group concerned *prediction of student performance*. Because all performance measures were at the school level and the sample size of nine schools was relatively small, no strong conclusions could be made. Nevertheless, some relations between the factors and student performance were determined.

In addition, some research limitations and future research directions were specified. The research limitations were related to the cross-sectional design of the study and the relative small sample size (despite the effort made to find participating schools). The future research directions were based on these research limitations and indicated that the focus for future research should be on the confirmation of the factors and their relations with performance measures.

The conclusion is that both parts of the project objective have been accomplished. The quality scan was scientifically tested and the practical relevance was encountered. Thus, the final conclusion is that the quality scan can be valuable for OCGH Advies.

1 Introduction

In this section, a context description of the project will be presented. This will consist of a description of the Primary Education in the Netherlands, the governmental evaluation of the Primary Education, and an introduction of OCGH Advies; the company that executes this project. Hereafter, the project objective will be formulated. In the last section, the report structure, the structure for the remainder of this report will be indicated.

1.1 Context description

The Primary Education has been a constant source of discussion for the Dutch government in the last couple of years. This indicates both the importance and difficulty of the realization of the Primary Education. The education is changing constantly. First, in the early nineties, it was considered very important that the content at schools was as diverse as possible. The students should learn the Dutch history, geography, Dutch language, mathematics, etcetera. The government prescribed a minimum level of knowledge that the students should acquire for each subject. Since the performance level for Dutch language and mathematics of students had decreased, elementary schools began focussing on these subjects again in 2003. Schools were autonomous in fulfilling their content. In addition, the government performed a monitoring role instead of an executive role. The government presented 58 "core targets". These core targets are handled strictly for Dutch language and mathematics, while schools have more freedom on how to present other courses. In addition, the government formulated a plan for Primary Education. The main contents are quality of education and education innovation, professionalism of teachers, relation between school, students and parents, and collaboration between school and social environment.[1] These aspects should improve the students' performance in Primary Education. Since the Primary Education is still performing below standard, there is still a long way to go (Reijn, 2007). In the newspaper article of Reijn (2007), it is indicated that students perform lower than should be expected on all subjects related to the Dutch language. Therefore, the government dedicates a lot of funds, up to €115 million in the next three years, for the improvement of Dutch language and mathematics. Thus, the last developments provide more freedom to the Primary Education of the realization of their lessons.

The Primary Education normally takes eight years. After finishing the Primary Education, the students advance to the Secondary Education. A class mostly has one or sometimes two part-time teachers per year. This can change each year. Almost all schools are associated to an above school board ("bovenschools bestuur"). These boards receive funds from the Dutch Government. The different schools present policy plans in which their strategies for a school year or a couple of school years have been described. In addition, the schools indicate how expensive their plans will probably be. Subsequently, the above school boards decide whether a school can have its budget or not. After this decision, the boards guard whether schools stick to their policy plans or not, otherwise the boards can decide to stop the financing of the plans. Furthermore, the above school boards try to support the schools with knowledge and courses.

1.1.1 Dutch Education Authority

As aforementioned, the Dutch Government performed a monitoring role in the Primary Education. This task is executed by the Dutch Education Authority ("Inspectie van het Onderwijs"). They guard the quality of all primary schools in the Netherlands. This needs to be done as objective as possible.

[1] Source: site Ministerie van Onderwijs; www.minocw.nl/onderwijs/index.html

Therefore, they use a standardized evaluation tool, which measures and evaluates the school on twelve different factors (Inspectie van het Onderwijs, 2005). The evaluation is done at least once per four years, but if a school is evaluated as insufficient, these evaluations are carried out more regularly. The twelve factors can be found in Table 1.1. Each factor consists of one or more indicators. A short description of the factors of the Dutch education Authority will be given here, and for a description of the underlying indicators, I refer to my literature study.

Quality management	Active role of students
Teaching program	Ambience at school
Instruction time	Accompaniment
Ambience in the class	Care
Quality of teaching	Performance of students
Tuning of education needs	Development of students

Table 1.1: Factors used by Dutch Education Authority.

First, *quality management* is the way the school determines, monitors, and improves the quality of education. An important aspect is the different educational needs of the students. The school should consider these differences when they formulate the content of their program. Further, the school should evaluate the results of students systematically and use this information to make improvements in the education program. Second, the Dutch Education Authority evaluates the *teaching program* of the school. An important aspect is that students can develop and prepare themselves for Secondary Education. Therefore, the indicators are mainly focused on the Dutch language and mathematics. Again, a distinction is made between the different educational needs of the students. Third, the *instruction time* is guarded by the Dutch Education Authority. It is important that students get enough time to learn the content. Fourth, it is measured whether teachers ensure that students feel supported in the class using the factor *ambience in the class*. Fifth, the *quality of teaching* is about the didactical actions of teachers. Teachers should encourage and support learning. Sixth, the factor *tuning of educational needs* measures whether the education program matches with the education level of students. Seventh, the *active role of students* is about the involvement of students in the educational program. The eighth factor is about the *ambience at the school*. This ambience should make sure that students and personnel feel safe. Further, it should support the parents' involvement with the school. The ninth factor guards the care the school takes for enough *accompaniments* of students. This should make sure that students develop themselves to their possibilities. Tenth, the factor *care* evaluates whether the school makes sure that students with a learning or development handicap get the appropriate care. This should be done by determining and evaluating the given care. The eleventh factor *performance of students* is an outcome measurement. It is measured to what extent the performance and skills of students at the end of the Primary Education is in line with the expected level. Finally, the factor *development of students* is more a process measurement. It measures whether the students develop themselves during the Primary Education as expected.

1.1.2 OCGH Advies

OCGH Advies is a consultancy agency for the Primary and Secondary schools, which is mainly active in the district Zuidoost Brabant. Since January 2008, OCGH Advies is a commercial organization (first it was (partly) subsidized by the Dutch Government). Before the commercialization, each agency had its own district it would work in, while due to this commercialization the agencies are free to approach schools in other regions as well. As a result, OCGH Advies has circa 50 competitors, since there are 50 advice agencies for the Primary Education in the Netherlands. However, each

agency is still mainly focused on their own region. Nevertheless, some agencies are trying to obtain projects at schools in the district of OCGH Advies.

OCGH Advies offers advice and support for the different parties of a school. This can be for the teachers, but also for the principal, the students, and their parents. The advice extends to four different areas. First, *education renewal* is focusing on new learning methods and instructional strategies for the teachers. Second, *student care* handles the care for students with a handicap, and the creation of a reasoned school-advice. Third, *organization development* takes care of the school's strategy. Not the individual teacher but the school is the focus here. Finally, *coaching and training* is meant for individual teachers. For instance, giving feedback, personal development, and coaching can help teachers who have problems in their class. In addition, OCGH Advies carries out projects for different Local Authorities ("gemeentes") and crèches.

OCGH Advies has about 50 employees working as (senior) advisors, in the back office, or as pedagogical psychological assistants. The latter are performing research at schools, which can be used by the advisors to formulate a grounded advice for the schools. In their goal of being a high performing partner for schools, the vision of OCGH Advies is formulated in four core values: reliability, quality, speed, and continuity. These values should ensure that the organization is focused on developing a high qualitative range of products and services.

1.2 Project objective

For my literature study, which preceded this Master Thesis project, I have studied the factors that possibly contribute to student achievement in Primary Education. Marzano's work (2003) was used as a starting point for this literature study. Marzano provided different categories at the school, teacher, and student level that can influence the performance of students. Additional literature was studied to confirm the categories presented by Marzano. This has led to thirteen categories that should have an influence on student achievement. Eleven of these categories were also formulated by Marzano (2003). Besides these, two additional categories, teacher motivation and leadership, were included, because these were found to be important in different studies (Butler, 2007; Pressley, Mohan, Raphael & Fingeret, 2007). The categories are presented in Table 1.2. In the second chapter, education aspects, the results of the literature study will be described in more depth along with its categories.

Guaranteed and viable curriculum	Classroom curriculum design
Challenging goals and effective feedback	Home environment
Parent and community involvement	Learned intelligence and background information
Safe and orderly environment	Student motivation
Collegiality and professionalism	Leadership (extra)
Instructional strategies	Teacher motivation (extra)
Classroom management	

Table 1.2: Categories influencing student achievement

In Marzano's book (2003), a questionnaire is provided with statements related to the categories. For the two extra categories, a literature search was conducted for instruments that can measure them. For measuring leadership a part of the Multifactor Leadership Questionnaire (MLQ) was chosen (Bass & Avolio, 2004). The statements of teacher motivation are based on a subscale of the Job Diagnostic Survey (Hackman & Oldham, 1980). As a result, the questionnaire that serves as a starting point for this project includes all the categories.

The questionnaire should be translated before it is applicable in the Netherlands. After this translation, it is still not clear however if all statements are formulated in such a way that they can be fully understood, what the relation between the statements is, whether they statistically form factors, and how exactly they are related to student achievement. OCGH Advies had the question whether it is possible to create a validated quality scan using a questionnaire adapted from Marzano (2003) and adjusted to the Dutch context, which has relations with school performance and other instruments. These other instruments are the standardized evaluation tool of the Dutch Education Authority and the EFQM model. The latter is a general used quality management model and will be described in more detail in chapter two. Therefore, the objective of my Master Thesis was formulated as follows:

Design and validate a quality scan instrument, which provides educational-related factors influencing student performance in the Primary Education.

1.3 Report structure

This study consists of different phases. These phases are organized into a model, presented in Figure 1.1. The arrows represent the relations between the different phases. This model provided a clear manual to start the study with. Additionally, these phases are used as a guideline for this report structure. First, phase one, literature review, is executed during the literature study. The main findings of this literature study will be summarized in chapter 2, *education aspects*. Hence, the different categories, that have influences on Primary Education, will be described and explained. In addition, a different instrument, the EFQM model, and the variables concerning student performance will be presented. In chapter 3, *research questions*, the research questions which should be answered during the project will be formulated. The analysis of these research questions will be executed during the phases two till six. Then, the results of the phases two, three and four will be presented in chapter 4, *designing the quality scan*. The translation method and process are indicated up to the version of the quality scan, which will be used for my analysis. Phases five and six will be described in chapter 5, *validating the quality scan*. The sample and measurements of the study will be presented. In addition, the different analyses will be formulated and the results of each analysis will be provided. Chapter 6, *implementation*, will provide the results of phase seven. The usability of the quality scan for OCGH Advies will be a central issue in this chapter. Finally, in the discussion, the main findings of the study will be presented and the research questions will be answered. Furthermore, the research limitations and directions for further research are given.

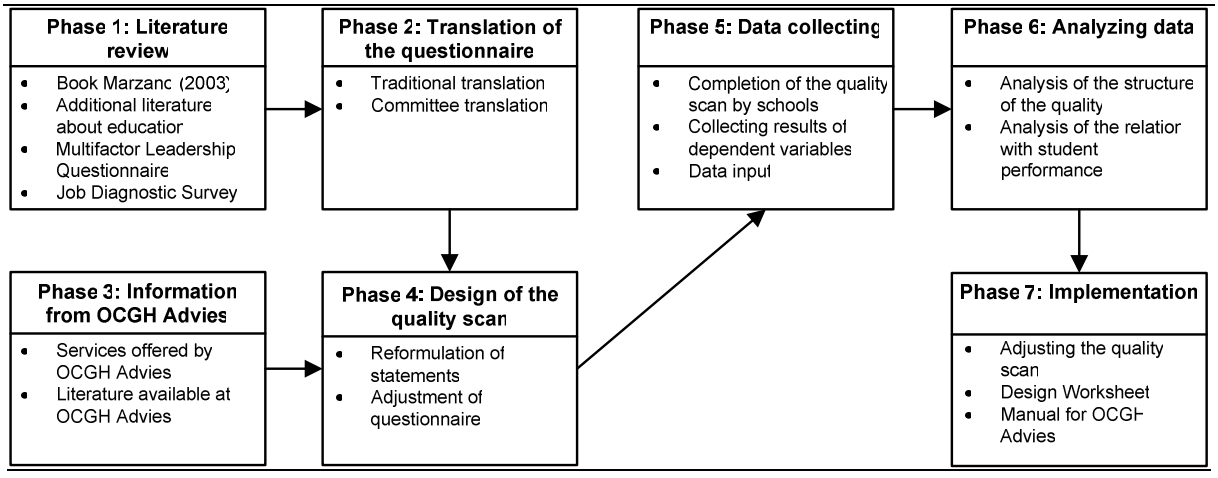


Figure 1.1: Research design model

2 Education aspects

Marzano (2003) conducted a meta-analysis study related to the Primary and Secondary Education. He used information of meta-studies performed by other researchers. This information was combined to indicate the effects of different categories on the performance of students. In addition, the strength of these effects was determined to understand the importance of each category. Each chapter of Marzano's book (2003) discussed one category. First, the importance of the category was specified by presenting the meta-analysis related to the category. Subsequently, the different aspects of the category were translated in action steps. These action steps were reformulated in statements, which ultimately were included in his questionnaire. Marzano found out that eleven different categories affect students' performance. He indicates that the categories can be divided into three different levels: the school level, the teacher level, and the student level. The results of Marzano's study show that the school level categories predict 6,7% of student performance. The teacher and student level categories predict respectively 13,3% and 80% of student performance. This specifies that the school and teacher can have a substantial contribution to the performance of a student. Nevertheless, it is important to verify these results with other research and field studies. Especially, since the conclusions of Marzano are mainly obtained by analyzing results of studies in the United States. Therefore, the importance of the different categories will be explained. In addition, the European Foundation for Quality Management model (EFQM model) will be presented since this model is regularly used at the Dutch primary schools. Finally, some measurements of student performance will be indicated.

2.1 School-level categories

In his book, Marzano (2003) mentioned five categories at the school level that influence student performance. Here, I will analyze each of these categories. The categories will be discussed in order of importance as indicated by Marzano. Besides, extra information in the literature related to these categories is presented.

2.1.1 Guaranteed and viable curriculum

Marzano (2003) indicates that the category *guaranteed and viable curriculum* consists of two parts. The *opportunity to learn* shows whether the students get enough chances to learn a specific content. A school provides an appropriate guaranteed curriculum if it takes into account the sequence of the contents, and checks if the specified curriculum is given. *Time* is the second part. The expected curriculum that a teacher should cover needs to be taught in the available amount of time. If there is not enough time available, the curriculum is not viable. Marzano defined time as the instruction time, because not all-available classroom time is instruction time. There will be interruptions, time for a joke, etcetera. Another definition of curriculum is formulated by Griffith (2008), namely curriculum is what schools are doing for students to achieve educational standards (e.g. the specific teaching material the school uses). These standards are, in concrete terms, the mission that a school should fulfil, and the curriculum should be derived from these educational standards. Moreover, the curriculum should consider different aspects, such as resource availability in school, learning style of students, and local environmental factors (Griffith, 2008). Furthermore, Fagan (1998) showed that different parties (i.e. teachers and parents) considered an available curriculum to be an important predictor of student performance. However, it is also suggested that it is not only the task of the government to create this curriculum, but this should be done in association with the teachers and parents (Fagan, 1998). The most optimal curriculum to arrive at is: a curriculum that takes into

account what subjects have to be taught, how these subjects need to be taught, and the time that needs to be available for each subject. This is in line with the two aspects indicated by Marzano (2003).

2.1.2 Challenging goals and effective feedback

The category *challenging goals and effective feedback* is focused on setting goals for and giving feedback at students (Marzano, 2003). In other words, the students should have high expectations or high pressure to achieve a goal. Close monitoring of the performance and development of the students is also necessary. With respect to goals, it is important that the goals are challenging for all students. Therefore, it can be useful to set both school goals and individual goals. Hence, a case study proved that goal-directed behaviour by students has a positive effect on the growth of mathematical skills (DiPerna, Lei & Reid, 2007). It is important to distinguish two different kind of goals used in an educational field, namely performance (or ability) goals and mastery (or learning) goals (Lemos, 1999; Lau & Nie, 2008). First, *performance goals* stimulate individuals to demonstrate high ability (Lemos, 1999). The feedback related to this type of goal, indicates whether a performance is achieved or not. Second, setting *mastery or learning goals* stimulate individuals to improve their competence (Lemos, 1999; Sideridis, 2005). Here, the feedback is focused on the performance improvement of a student. Thus, it is not argued that all students should gain a specific result, but the students should enhance their performance in comparison with their earlier performance. Lau and Nie (2008) found that, at both the classroom level and the personal level, the mastery goals for mathematics were related to adaptive outcomes, such as student engagement and interest. Performance goals will not always have a positive effect on student performance, since below average students will get negative feedback, which creates negative affective reactions. Therefore, their performance will decrease (Sideridis, 2005).

The feedback offered to the students needs to be within a specified timeframe. The best moment to provide feedback to a student is during the learning of the content. This is called formative feedback and is especially useful for development and learning (London & Sessa, 2006). If the feedback is given during the learning of the content, the student can change his/her effort and increase his/her performance. A related characteristic is that the feedback should make a comparison with historical performance of the student and not compare the student with other students, because the student can influence his/her own performance but not the performance of others (Sideridis, 2005). Another desirable characteristic of feedback is that it has to be specific to the content being learned. Besides, if the feedback is not confidential, the teacher should watch for the creation of social goals. Social goals are goals set to satisfy classmates (Spera & Wentzel, 2003). Since students want to be part of the group and do not want to be teased by their classmates, this may have negative consequences such as less concentration on learning goals and setting lower learning goals (Spera & Wentzel, 2003).

2.1.3 Parent and community involvement

This category is about the amount of involvement and support of parents and the community with the school. Marzano (2003) argues that three features define the effectiveness of the relation between the parents and the community on one hand, and the school on the other. First, communication between the parties results in a clear view of each other's opinions. Epstein (2005) indicates that teachers lack the knowledge to work effectively with parents and community partnerships. This means that the involvement of the parents and community is not utilized completely, although it could increase the students' performance. The second feature is participation of the parents, i.e. the involvement in the day-to-day running of a school, for example a guest lecture. Englund, Luckner, Whaley and Egeland (2004) studied the effects of parental involvement and expectations with a longitudinal design. They

found that both aspects affect student achievement. Finally, school governance can improve the effectiveness of the parent and community involvement. It stands for having a clear structure in decision making with respect to the parent and community's voice. However, the importance of parent involvement is concluded in other studies, but the importance of community involvement is less clear. Just church involvement is mentioned in some studies. Therefore, the community involvement should perhaps not have to be included in the quality scan.

2.1.4 Safe and orderly environment

Marzano (2003) specifies *safe and orderly environment* as the fourth category at the school level. It is argued that students and teachers need to feel safe to learn and teach well. Without a minimum level of safety and order, the school cannot have a positive effect on student performance. In addition, if clear rules are presented violence and truancy will also decrease. Koth, Bradshaw and Leaf (2007) studied the underlying aspects that predict the students' perception of safety at elementary schools. Their results showed that the school's climate explains up to 27 percent of the variance of order and discipline. However, student level accounted for more variance (65 percent). Another study points out that security and maintenance of the building had no effect on student performance (Schmitt, Sacco, Ramey, Ramey & Chan, 1999). According to Schmitt et al. (1999), Primary Education students are more sensitive to classroom climate than school climate. As a results, it is highly uncertain if there will be a relation between this category and student performance. Nevertheless, safe and orderly environment will be included in the quality scan.

2.1.5 Collegiality and professionalism

For the fifth category, Marzano (2003) used the term *collegiality and professionalism*. This category presents the interaction among staff members and the professional approach of teachers. The interaction refers to sharing experiences, demonstrating respect for each other, and analyzing and critically evaluating practices and procedures. An aspect of professionalism is the efficacy of teachers. In the educational context, teachers' self-efficacy is formulated as the teachers' belief in their ability to influence students' outcomes and accomplish instructional goals unrelated to the environmental aspects (Wolters & Daugherty, 2007). An important outcome of their study is that self-efficacy was related to the mastery goals that were set in the class, which has also a direct effect on student achievement. Another aspect is the pedagogical knowledge of a teacher. Pedagogical knowledge is the knowledge about the way a specific content should be taught.

2.2 Teacher-level categories

The categories related to the teacher level are the most important categories for this study. As mentioned earlier, Marzano (2003) revealed that these categories predict 13,3 percent of the variance in student achievement. This is almost twice as much as the school level categories. Thereby, the school and especially its teachers can still influence the teacher level categories. In this section, the three categories related to this level will be presented and compared with other studies.

2.2.1 Instructional strategies

The teacher's instructional strategies influence the students' achievement. Teachers that master more instructional strategies are more effective teachers than teachers that are only proficient in a few instructional strategies (Marzano, 2003). There are many instructional strategies, but a teacher does not have to be good at all instructional strategies. The instructional strategies can be divided in three categories; strategies used for regular unit intervals, strategies used for input experiences, and strategies used for reviewing, practicing, and applying content. A teacher should master at least some

instructional strategies in each category. The strategies used for regular unit intervals focus on setting clear learning goals. These goals should be set at the beginning of a lesson and students must have the possibility to set their own learning goals. With the strategies used for input experience, teachers provide students with contribution regarding a unit's content. Finally, due to reviewing, practicing, and applying content, teachers give students the opportunity to make changes, additions, and corrections to their initial understanding of the content. The literature provides a large amount evidence for the importance of different instructional strategies. Since a full explanation of all instructional strategies is beyond the scope of this report, only two important conclusions are presented. First, as aforementioned, not all strategies need to be used by the teachers. However, effective teachers will use more strategies than ineffective teachers (Reeve & Jang, 2006). Second, the teacher should use diverse instructional strategies on students with different abilities, since an instructional strategy has not the same effect on the performance of each student (Lapadat, 2002).

2.2.2 Classroom management

Marzano divided *classroom management* in four different areas. First, the teacher should establish and enforce *rules and procedures*. These rules and procedures can differ per classroom, but without these rules and procedures, the students have no guidelines on how to behave. This makes effective teaching impossible. Second, the teacher should carry out *disciplinary actions*. This is most effective if a teacher uses both reinforcement, using rewards to stimulate good and to end negative behaviour timely, and punishment for inappropriate behaviour. This is remarkable, since research mainly argues that punishment is ineffective (Butterfield, Trevino & Ball, 1996). However, Butterfield et al. (1996) indicated that many managers also use punishment. It can be effective if the employees are expecting to be punished. This can also be the case with students. If students are doing something wrong, they and their classmates expect them to be punished. Third, the teacher should conserve an effective *teacher student relationship*. This means that a teacher should be neither too dominant nor too cooperative. If the relation is good, the student will listen better to the teacher, and will accept the rules and procedures set by the teacher. This is inline with the study of Hughes, Luo, Kwok and Loyd (2008). In a case study executed in Texas, they analyzed that students who experience a quality teacher-student relationship, a relationship that promotes warmth, support, and low levels of conflict gain more achievement. Finally, the teacher should maintain an appropriate *mental set*. This consists of two features "withitness", and emotional objectivity. Withitness is defined as "knowing when to intervene before behaviour will become inappropriate" (Marzano, 2003, p. 94). An emotional objective teacher will stay calm instead of getting angry or upset when students violate rules and procedures. In addition, Babad, Avni-Babad and Rosenthal (2003) concluded that a teacher's non-verbal behaviour influences students' perception of the teacher. Thus, the teacher should not react emotionally in specific situations (not in verbal or non-verbal way); because it influences the feelings students have about their teacher.

2.2.3 Classroom curriculum design

Classroom curriculum design is defined as the sequencing and pacing of content along with the experience students have with that content (Marzano, 2003, p. 106). Most teachers do not make an explicit decision on how the sequence and pace of contents should be, but they use the sequence provided by the lecture books. Therefore, teachers often do not take into account the needs of their students and three important principles of learning. First, the specific knowledge that is the focus of the lesson should be identified, before learning of that lesson can be enhanced. This means that the teacher should think about this knowledge and should not just use the lessons out of a lecture book. Subsequently, if students recognize the similarity between contents, this will enhance their learning.

Thus, teachers should indicate if a lesson is a logical successor to a previous lesson. Finally, students should have multiple exposures to and complex interactions with knowledge, before it enhances learning. In contrast with instructional strategies, there is little research available about the effect of classroom curriculum design. Nevertheless, this subject is important, because good instructional strategies can only be effective if the right content has been taught. One study of Li (2002) reveals this importance. Many tests showed that Chinese students excel on mathematics compared to the American students. The reason for this is that Chinese teachers put more effort in the curriculum for their class. The Chinese teachers do not only have a good understanding of the content they are teaching, but also have a clear understanding of the sequencing of the content. Thus, the importance of good classroom curriculum design is proved, but not in so many studies.

2.2.4 Teacher motivation

This category is not indicated by Marzano. However, multiple studies showed the importance of teacher motivation for the student enhancement (Roth, Assor, Kanat & Kaplan, 2007; Butler, 2007). Roth et al. (2007) tested the effect of motivation at elementary schools in Israel with a comparison between autonomous motivation and controlled motivation. Autonomous motivation enables people to realize their authentic self. At the other end, there are people with controlled motivation; they experience the motivation as external and internal pressure (Ryan & Deci, 2000). Their conclusions are that teachers who have autonomous motivation experience higher personal accomplishment and feel less exhausted. In addition, these teachers provide lessons that support autonomy. This means that teachers' behaviour is focused on the mastery goals of students. The students like to learn, because they see it as valuable and interesting. Finally, teachers' autonomous motivation promotes students to motivate their learning autonomously (Butler, 2007).

2.3 Student-level categories

Marzano's final categories are at the student level. As mentioned earlier, the student level contributes to the highest variance in student performance. In this section, it will be indicated whether the teacher or school can influence these categories and how this can be done.

2.3.1 Home environment

The category *home environment* consists of three elements, *communication about school*, *supervision*, and *parental expectations and parenting styles* (Marzano, 2003). First, communication is about the parents' interest in and communication about the schoolwork of their children. Trautwein and Lüdtke (2007) showed that the communication between parent and student on school-related subjects has a positive effect on the homework effort of the student. Therefore, they concluded that good communication motivates the student to put effort in the task. Second, supervision is about the parents' control and monitoring of their children. This aspect is less clearly related with student achievement, because it has both a positive and negative relationship. Students with low performance are monitored more intensely by their parents. Finally, parental expectations and parenting styles is the aspect strongest related to student achievement. Marzano (2003) indicates that an authoritative parenting style (rules with input of children and without negative emotion), is positively related to student achievement. The two other styles, authoritarian (many rules, without children's input) and permissive (no rules) are negatively related to student achievement. A school can provide training for the parents, where this importance can be explained and advice can be presented.

2.3.2 Learned intelligence and background knowledge

Learned (or crystallized) intelligence characterizes intelligence as knowledge that can be learned. This type of intelligence depends on experience and fluid intelligence, which characterizes intelligence as a cognitive process (Marzano, 2003). The difference between crystallized intelligence and background knowledge is that crystallized intelligence is learned knowledge about the world and background knowledge is learned knowledge about a specific domain. However, for predicting student achievement crystallized intelligence and background knowledge can be seen as identical. A good indication of these two constructs is a student's vocabulary knowledge. Especially in the lower classes of the elementary school, the vocabulary knowledge is a predictor of future student achievement (Biemiller & Boote, 2006). Biemiller and Boote (2006) found that students can learn 40% of the words taught by direct teaching of the word's meanings. Thus, with a good teaching program related to vocabulary, the school can enhance this category.

2.3.3 Student motivation

Marzano's (2003) last category on the student level is *student motivation*. In general, motivation consists of three aspects (1) what energizes behaviour, (2) what directs such behaviour, and (3) how the behaviour is sustained (Porter, Bigley & Steers, 2003). Trautwein and Ludtke (2007) proved that the self-reported effort of homework has a positive relation with student performance. Besides, they showed that students who are putting more effort in the contents are scoring high on student motivation. Reeve and Jang (2006) indicated that the quality of the teacher-student relationship affects student motivation. This means that teachers influence a student's motivation and that this influence can to some extend be controlled by the teacher. For instance, providing accurate feedback regarding self-efficacy, providing stimulating and interesting tasks, using evaluation structures that promote mastery goals, and providing opportunities to exercise some control by the students enhance student motivation (Pintrich, 2003). A teacher can have therefore influence on the student motivation by providing a stimulating context.

2.4 Leadership

Marzano (2003) argued that adequate leadership of the principal is related to all categories presented in his study. If the principal provided a clear vision and would stimulate his or her school team and students, this could enhance the performance at a school. Marzano (2003) did not include leadership in his questionnaire since the affect of leadership is difficult to determine. Still, a study by Pressley et al. (2007) showed that a new principal had a positive effect on student achievement in an elementary school in the US. The year before she arrived, 38% of the students passed a writing test, while a year later, 84% of the students were passing the writing test. This was suggested to be a result of her vision about and strategy regarding students' reading performance. The study of Pressley et al. (2007) showed that leadership is related to student achievement and that it is possible to measure the effects of good leadership on student achievement. In addition, a different study indicated the effectiveness of leadership at a school in the Netherlands (Geijsel, Sleegers, Van Den Berg, 1999). Thus, a measure of leadership should be implemented in the quality scan. The exact implementation will be explained in the chapter design methods.

2.5 EFQM model

A general model used to conduct a self-assessment by companies is the European Foundation for Quality Management model (EFQM model). This model is useful for all kinds of companies and it is a model generally used by Dutch schools to conduct a self-assessment. Since the Dutch schools are familiar with the EFQM model, it can be useful to compare the EFQM model and the quality scan.

The goal of the quality scan is to provide a complete picture of the quality of several aspects at a school. The EFQM model can also be used to indicate the completeness of the quality scan since the EFQM model also analysed the quality of an organization or school, and the practical relevance of this model has been established in several contexts,. Furthermore, the EFQM model divides the factors that determine the quality of an organization into several area of attentions, thus it can be used to classify categories of the quality scan to these criteria.

A graphical presentation of the EFQM model is presented in Figure 2.1. The model consists of nine criteria. Five of these criteria can be manipulated by the organization, these are called the "enablers", and the other four criteria represent what the organization can achieve, called "results" (Wongrassamee, Gardiner and Simmons, 2003). Displaying the connection between the enablers and the results is the strength of this model. Each criterion reveals a separate part of the organization and together they present the total quality of an organization (Nederlandse Kwaliteit, 1994). Here, a short description of each criterion is presented, based on Wongrassamee et al. (2003) and Nederlandse Kwaliteit (1994). First, *leadership* represents the behaviour of the managers, the way they inspire their organization, and how they try it to continuous improve. Second, *people management* is the way the organization handles its employees and how it develops the full potential of its employees to increase the business processes. Third, *policy and strategy* deal with the organization's value, vision, and strategic direction plus how this will all be accomplished. Fourth, *resources* describe how the organization manages its external partnerships and internal resources effectively. Fifth, the criterion *processes* indicates the way the organization controls its activities and processes. Subsequently, the criteria for the results should be explained. The first three criteria are *people satisfaction*, *customer satisfaction*, and *impact on society*. These criteria indicate if the organization fulfils the needs and expectations of its personnel, its customers, and the society respectively. Finally, the *business results* evaluates whether the planned business performance is accomplished and the needs of the shareholders are satisfied.

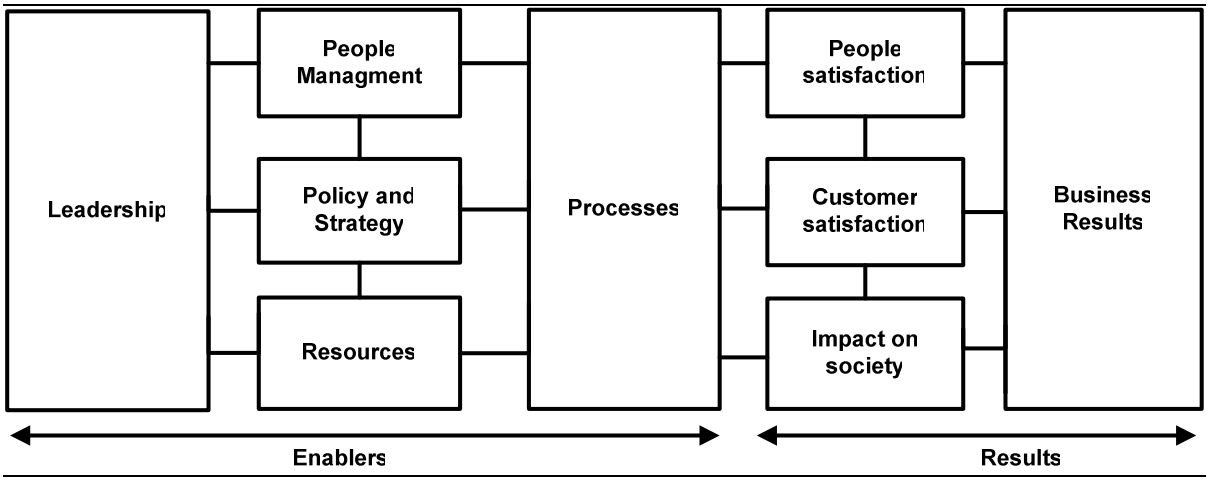


Figure 2.1: The EFQM model of Excellence (Ehrlich, 2006)

2.6 Student performance

An analysis between the quality scan and some student performance measures will be presented in the result section. Therefore, it is necessary to give a description of these student performance measurements. There are three requirements related to these variables. First, for an accurate comparison between schools, it is important that the dependent variables can be collected from (almost) all schools. This will enhance the sample size of the different analyses, which increase the reliability of the conclusions. Second, the relation between the independent variables and dependent

variable should be defined. For example, when the relation between the teacher level categories and the dependent variable is analyzed, an appropriate dependent variable should be chosen. It is important that the effect of other variables on the dependent variable is minimized. For example, not only a teacher who teaches a class this year, but also the teachers who taught the class before can have an influence on the class' performance. A choice should be made whether to take the current results of a test or the progression of the results since the previous test. For every analysis, these choices need to be made. The third requirement is the time aspect. The time span between the realisation of the quality scan and the time that the student performance is measured should be as small as possible.

The above described requirements have led to three different student performance measurements. The first measurement is the Cito exam at the end of the Primary Education ("Cito-eindtoets"). This exam is used by 6300 primary schools in the Netherlands (Terugblik en resultaten, 2008). The Cito exam uses a standard score ranging from 501 to 550. The mean and standard deviation of the exams in the Netherlands are known. Therefore, it is not only possible to use the standard score of a school, but also the performance with regard to the mean of all schools in the Netherlands. Second, some tests of the student monitoring system ("leerlingvolgsystem") are used. These tests measure the mathematics or reading skills of students and can be carried out in all eight years. This dependent variable is useful to measure the relation between the teacher level categories and a class. Finally, it is collected how many students went to the different levels of the Secondary Education in the previous two years. This indicates the student performance at the end of the Primary Education.

3 Research questions

In the previous chapter, the main findings of the literature study were provided. During the project, I will not test specific hypotheses, since the Master Thesis project is mainly design focused; the goal is to design an appropriate quality scan. Still, some empirical research questions are formulated that will be answered in the results and discussion sections. These research questions are partially based on findings that are encountered during my literature study and partially out of interest in specific relations. They are divided in different subjects and are presented in this chapter.

3.1 Generalisation

As mentioned earlier, Marzano's questionnaire (2003) was developed for the American education. This means that the questionnaire is not automatically applicable for the Dutch Primary Education. Thus, it is possible that some adaptations in the formulation of the questionnaire are necessary before it is applicable in the Netherlands. This has led to the first two research questions.

1. *Is it possible to use the quality scan for the Dutch Primary Education?*

2. *What adaptations - if any - have to be made to make the questionnaire formulated by Marzano (2003) suitable for Dutch Primary Education?*

The first research question has partly been handled during my literature study. Additional literature showed that the categories indicated by Marzano (2003) were effective in other countries as well. This applied not only to the Netherlands, but also for countries such as Israel and China. Furthermore, it is concluded that the categories at all three levels can influence student performance in the Netherlands. Besides, the Dutch Education Authority measures the Primary Education with objective factors and underlying indicators (Inspectie van het Onderwijs, 2005), which indicates that different parts of a Dutch primary school can be evaluated separately. Nevertheless, the additional literature also showed that aspects of Marzano's questionnaire were not complete. There are two extra categories that can have influence on student performance; leadership and teacher motivation (Pressley et al., 2007; Roth et al., 2007). Since these two categories are not implemented in Marzano's questionnaire, additional items are necessary to measure these categories. The implementation of these two categories is the first adaptation of the questionnaire by Marzano (2003). In addition, this adaptation results in the following research question:

3. *How should teacher motivation and leadership be included in the instrument?*

This research question is also already analyzed during my literature study. For both categories, different instruments have been examined. As indicated in the introduction, a part of the Multifactor Leadership Questionnaire (MLQ) is used for the category leadership (Bass & Avolio, 2004). The MLQ consists of two different types of leadership, namely transformational leadership and transactional leadership. A transformational leader motivates subordinates to do more than they originally thought was possible (Bass & Avolio, 2004). The transformational leader is a source of inspiration, a source that elevates the needs of the subordinates, presents new perspectives about the world, and is trusted by the associates. A transactional leader recognizes and clarifies the roles and tasks associates require to reach the desired outcomes (Bass & Avolio, 2004). This type of leadership can lead to an increase in the effort associates put into their work. However, most of the time, this provides a non-significant change. Both Geijsel et al. (1999) and Avolio (1999) proved that in schools

transformational leadership is more effective than transactional leadership. Therefore, only the statements related to transformational leadership were used in the questionnaire. Nevertheless, choices had to be made about which statements to include in the quality scan, because due to the length of the quality scan not all statements could be included. The transformational leadership statements are divided in five issues, each issue consists of four statements. The issues are; attributed idealized influence, behaviour idealized influence, inspirational motivation, intellectual stimulation, and individualized consideration (Bass & Avolio, 2004). Since ten statements about leadership would be an appropriate number for the length of the quality scan, two statements per issue could be included. Three people, a professor of the Eindhoven University of Technology, an advisor of OCGH Advies, and I have chosen the statements we believed to be most applicable for an effective principal of a primary school. The statements of teacher motivation are partly based on the internal work motivation measure of Job Diagnostic Survey (JDS) of Hackman and Oldham (1980). In addition, Steehouwer (2007) developed four questions to measure motivation for his Master Thesis. Again, the most appropriate statements were chosen from these two sources for the Primary Education. This resulted in four statements related to the category teacher motivation.

3.2 Properties of the quality scan

In the previous section, research questions were formulated that should result in a complete quality scan. Nevertheless, the properties of the quality scan are not yet clear. The different categories consist of a number of items. It is of importance to indicate whether these categories are collective terms of the items, or real statistical factors that are a combination of the items in the quality scan. This distinction provides insight on how the results of the tool should be interpreted. Besides, it is possible that there are moderate correlations between the categories. For example, a relation between student motivation and teacher motivation (Roth et al., 2007), or the relation between instructional strategies and classroom curriculum design (Li, 2002). When these relations are indicated, it is possible to analyse the current situation at a school even better. Even though it is not yet clear whether there is a relation for every combination of categories, it is interesting to analyze such relations. These issues result in the next research questions:

4. *What is the reliability of the categories and factors used in the quality scan?*

5. *What is the underlying structure (correlation) among the factors included in the quality scan, and what are the consequences for the improvement plans?*

3.3 Relation with other instruments

Similar to most practical projects, the current quality scan will not be the only instrument in place. If the relationship between the quality scan and the other instruments is clear, the usefulness of the quality scan will increase. The two other instruments used in the Dutch Primary Education are the EFQM-model and the instrument used by the Dutch Education Authority. The EFQM-model is generally used by the Dutch primary schools for indicating the quality of the schools. Thus, a clear relation between the criteria of the EFQM-model and the items or factors of the quality scan would enhance the confidence in and commitment to the quality scan. In addition, the instrument of the Dutch Education Authority is used to provide an objective evaluation about the quality of the school in the Netherlands. The quality scan shows the strengths and weaknesses of the school. Therefore, by using the quality scan, schools can improve themselves before the Dutch Education Authority measures their performances, which clearly is an advantage. These possible relationships result in two research questions:

6. *Can the statements or factors be ascribed to the different criteria of the EFQM model?*

7. *Is there a relation between the statements and/or factors of the quality scan and the indicators of the Dutch Education Authority?*

3.4 Prediction of student performance

In the previous section, the importance of the relation to other instruments was discussed. Nevertheless, the relation between the quality scan and the students' performance at the school is the most important issue. It is preferable that the quality scan is related to objective measurements of student performance. The relation between a factor and the student performance can both be direct and indirect. For example, the school related categories can influence student performance, but they can also influence teacher level categories, which are related to student performance as well. Moreover, Marzano's study (2003) shows that the school, teacher, and student level categories predict respectively 6,7 percent, 13,3 percent, and 80 percent of student performance. This study will be carried out at different schools and with multiple teachers per school. Thus, with a regression analysis it will be possible to indicate the percentage of the student performance variance that is caused by different factors. For each factor, it is known to which level it belongs. Still, not all factors related to the student level, for example intelligence, will be implemented in the quality scan. Thus, the quality scan will not predict all the variance. The part that is not predicted by the quality scan can be caused by factors related to student level. Therefore, I formulated the next research questions:

8. *What is the relation between the factors of the quality scan and quantitative indicators and measurements of educational effectiveness at the school?*

9. *How much student performance variance is caused by the school, teacher, and student level factors?*

4 Designing the quality scan

In this chapter, the design of the quality scan will be described. The methods used for the design of the quality scan will be presented in the first section. The remaining sections of this chapter will provide the results of the design process. The goal of designing the quality scan was to use the final version to validate the quality scan with a field study.

4.1 Design methods

The design of the quality scan was executed in different steps. These steps should have provided a structured and orderly design process. They guaranteed that the design process was clear and accurate. Each of these steps will be described separately in this section.

4.1.1 Translation

As mentioned earlier, Marzano's questionnaire (2003), and parts of the Multifactor Leadership Questionnaire (Bass & Avolio, 2004) and the Job Diagnostic Survey (Hackman & Oldham, 1980) were used as a starting point for the quality scan. Since the statements for all three questionnaires were in English, these needed to be translated into Dutch first. This translation had to be done as accurate as possible since it provided a first version of the Dutch statements. Therefore, different steps were executed during the translation process. Balbinotti, Benetti and Terra (2006) present different methods for the translation process. There were two methods used during this translation process. First, *traditional translation*; one person translated the original version. With this technique, there is a high possibility of bias in the translation, because of language barriers and knowledge biases by the researcher. Therefore, a second technique, called the *committee translation,* was also used during the translation process. Here, a committee or translation group discussed the translation of the researcher (Balbinotti et al., 2006). This translation group consisted of a professor of the Eindhoven University of Technology and me. In addition, a sworn translator evaluated the translation.

4.1.2 Reformulation

The quality scan will not be an exact translation of the original version, since this version was developed for American education. Therefore, a couple of adjustments had to be made. This way, the quality scan would become more accurate for the Dutch education. These adjustments were based on the input of an advisor of OCGH Advies, who is familiar with the Dutch education and its teachers. The suggestions of the advisor were discussed in a committee, consisting of the advisor, the professor of the Eindhoven University of Technology and me. The goal of the committee was to guard that the original statements would not change too much and that the wording of extra statements were in line with the original statements.

4.1.3 Structure of the quality scan

During the design process, the structure of the quality scan should also be designed. The previous two steps generated accurate statements, while this step had to make sure that these statements should be structured in a clarifying questionnaire. The same course of action as with the translation process has been followed for this process. First, I structured the statements and designed the lay-out for the questionnaire. Afterwards it was discussed in a committee. Again, this committee consisted of the advisor, the professor, and me.

4.1.4 Pilot test

The next step of the design of the quality scan was testing the scan in practice. The quality scan was in first instance filled in by teachers of one school. These teachers could give criticism on the quality scan, which could be analyzed hereafter. A statistical validation of the pilot version of the quality scan was not possible, since the sample size was too small (N=18). Nevertheless, with specific comments of the teachers, it was able to determine which statements needed to be adjusted and what kind of adjustments were necessary.

4.1.5 Final adjustments

Finally, together with the necessary adjustments based on the pilot test, additional information about the quality of the content of the questionnaire was gathered. This was done by interviewing two advisors of OCGH Advies and a second professor of the Eindhoven University of Technology. Besides, literature of teachers' education was used, since this provided general used terms by teachers. These two actions led to the final experimental version.

4.2 Step 1: Translation

The previous section provided the design methods used. In this section, the results of the first step will be presented. As described earlier, The English statements of Marzano's questionnaire (2003), the Multifactor Leadership Questionnaire (Bass & Avolio, 2004), and the Job Diagnostic Survey (Hackman & Oldham, 1980) were first translated into Dutch. This was done by using a dictionary for the specific terms. This was the traditional translation technique, which was followed by the committee translation technique. The professor received the translation, checked this version, and afterwards we discussed his suggestions. These suggestions were mainly related to the structure of the phrases. I then adjusted and rechecked the statements. Subsequently, a sworn translator reviewed the statements. Again the structure of the statements was improved, especially because the correct conjunctions were included. This translator also formulated some terms of the statements less formal. For example, the term *identified* was first translated in *geïdentificeerd* and became *vastgesteld*. A second example is *non-linguistic*, which initially became *niet schriftelijk* and was improved to *non-verbaal*. These remarks were once again discussed by the professor and me. This resulted in an exact translation of the statements into a correct Dutch sentence.

4.3 Step 2: Reformulation

During the second step, the exact translations of the previous step were analyzed and reformulated so they became more applicable in the Netherlands. As mentioned earlier, this was done by an advisor of OCGH Advies and afterwards discussed by the advisor, the professor, and me.

First, it was indicated that some statements were not useful in the Dutch Primary Education. One of the reasons that statements would not give extra information is that the content of the statements is required by Dutch law. As a result thereof teachers are obliged to do certain things and all Dutch schools will be executing the action(s) put down in the law. For example, each school should have a parent council ("medezeggenschapsraad"), which makes statements concerning the presence about parent involvement in the governance of the school superfluous. A second reason for deleting statements was that they exactly measured the same thing as another statement. The last reason for removing statements was that they were too vague. It was not possible to have an unambiguous perception of the statements. This resulted in the removal of seven statements, which were originally included in Marzano's questionnaire.

Second, a couple of statements were allocated to a different category. A first source of reallocating a statement was the changed content after the statement was rewritten. For example, the content of one statement was changed from the detection of violence and extreme behaviour into the detection of students with learning problems, since the first is not an issue in the Primary Education while the latter is an important problem. As a result, the statement matched more with the category *challenging goals and effective feedback* than with *safe and orderly environment* in which the statement was categorized originally. A second reason for reallocating statements was to remove unnecessary categories. The category *home environment* had just one statement and with the reallocation of the statements this category did not need to be implemented in the quality scan. This made the quality scan more compact. The second category that was removed is *learned intelligence and background knowledge*. The English name for this category is vague and there was no appropriate Dutch noun for this category.

Third, there were statements in the quality scan that had multiple meanings. Double meanings make it harder for a respondent to give a clear answer to the statement. For instance when one part is executed at the school while the second part is not executed at the school. Such statements were split up in two or more statements. For example, the statement concerning the use of individual and school goals for future actions was split up. With the separation of the individual and school goals, a respondent can indicate whether one is used or not and if the other is used or not. This separation of statements led to five additional statements.

Finally, the advisor of OCGH Advies indicated that there were some subjects relevant in the Dutch Primary Education, besides leadership and teacher motivation, which were not available in the questionnaire of Marzano (2003). This came to my attention when it was tried to indicate relations between the indicators of the Dutch Education Authority and the statements of the questionnaire. This was independently done by three persons. Later, the relationships were compared. There were a lot of differences between the results of the three persons. Nevertheless, for different indicators and even factors of the Dutch Education Authority no one had a related statement of the questionnaire. The content of most of these indicators concerned the differentiation of activities for students with different knowledge levels. The importance of the differentiation of activities also followed from the literature, since it has been indicated that not all instructional strategies have the same effect on the performance of each individual student (Lapadat, 2002). Thus, it is beneficial to provide some students with different activities, as to let them understand the content more accurately. Therefore, three statements about differentiation were implemented at the category *classroom curriculum design*. Second, the collaboration of teachers was not implemented in the questionnaire. However, if teachers stand up for themselves and are critical to each other, this can increase their teaching skills and knowledge (Lovelace, Shapiro and Weingart, 2001). Therefore, the category *collegiality and professionalism* was extended with three extra statements.

4.4 Step 3: Structure of the quality scan

The first two steps both focused on the statements of the quality scan. However, an appropriate structure of the questionnaire was also important. First, it was necessary to have a clear introduction. That way the respondents were motivated to complete the questionnaire. Therefore, the intention of the questionnaire was described, and it was presented that it concerned the opinion of the respondent and therefore there were no wrong answers. Additionally, it was explained that it would take 30 minutes to fill in the questionnaire.

Besides, some general information was asked before the real statements were presented. This general information included the name of the school, the gender of the respondent, years of employment at the school, the respondent function, and the group the respondent is teaching. This information could be useful for some extra analysis of the data.

Hereafter an appropriate scale had to be chosen. Marzano (2003) used a four-points scale, with three different questions for each statement. These questions are (Marzano, 2003, p. 160 and 163):

1. To what extent do we engage in this behaviour or address this issue?
2. How much will a change in our practices on this term increase the academic achievement of our students?
3. How much effort will it take to significantly change our practices regarding this issue?

The goal of this quality scan was to determine the current situation at a school. Therefore, the first question was sufficient. The other two questions were indicating the effect of possible improvement steps. These questions can better be handled during a meeting following after the results have been shown. The first question was translated to different terms per scale point. Besides, a five-point Likert Scale was used instead of a four-point scale. This gave a higher chance for significant differences between the answers of respondents. The scale anchors were numbered one to five. Scale anchor one was indicated by the concept *(bijna) nooit* ((almost) never) and scale anchor five by the concept *(bijna) altijd* ((almost) always). The scale anchors in the middle had no concept name, since there were no concepts available which indicated similar distances between the scale anchors (during the pilot test these had concept names).

Finally, below each category label, a short description of the category was presented. With these descriptions, the meaning of the categories should be better understandable for the respondents.

4.5 Step 4: Pilot test

After the first three steps the quality scan was ready for a pilot test. The goal of the pilot test was indicating whether all statements were clear for the respondents. This pilot version of the quality scan was completed at one school, which provided eighteen respondents. Before the respondents received the questionnaire; it was explained that they could include remarks for the statements. In general, the following comments were made:

- Some statements were too abstract. The respondents found it difficult to understand the meaning, due to the use of specific terms.
- The respondents had difficulties answering the statements at the teacher level, since these should be answered for the teachers in general at the school.
- For the category *leadership*, the respondents were afraid that their answers would be used against them by the principal. This would lead to socially desirable answers.

In the next section, these remarks will be handled for creating the final experimental version of the quality scan.

4.6 Step 5: Final Adjustments

As indicated in the previous section, the respondents in the pilot test provided some remarks, which were analyzed. Besides, two additional advisors of OCGH Advies and a professor of the Eindhoven University of Technology checked the quality scan. They had some suggestions to improve the quality scan. These suggestions were also considered and implemented in the quality scan. In this section, the adjustments, based on the remarks and suggestions, are presented.

During this step different statements were once more changed. There were different reasons for changing the statements. First, some statements were too abstract, this could be improved when combined with the experience of the teachers. For this purpose, a book of the teacher education program was studied. Leenders, Meyer, Sanders and Veenman (1993) describe effective instruction forms for the Primary Education. The terms used in their book were used for reformulating the statements. This way the concepts would be more familiar for the teachers. For example, *activiteitenaanbod* is replaced by *werkvormen* and *erkenning* is changed in *successen*. A second reason was that some statements were too long. A professor at the Eindhoven University of Technology argued that short statements that are to the point were the best kind of statements. Therefore, some statements were rewritten with the objective of abbreviating them. Third, the respondents of the pilot test commented that they had difficulties answering the statements, because the statement should be answered for all teachers (i.e. "The teachers in my school…"). This was also pointed out by an advisor of OCGH Advies. He argued that this would create social desirable results. Therefore, the statements at the teacher level and some statements of *collegiality and professionalism* were rewritten in the first person (i.e. "I…"). This adjustment made the statements more personal for the respondents. It is argued in the literature that this change will not have large consequences for the aggregation of the data (Klein, Conn, Smith & Sorra, 2001). Later in this report, this issue will be described in more detail. Fourth, statements were split up or a term was removed, because of the double meaning of the statement. An example of splitting a statement in two statements was the statements which concerned actions related to classroom and homework exercises. The classroom and homework exercises were separated into two statements. An example of the extraction of a term from a statement is shown by the statement concerning providing training and help for the parents. Here, the term *training* was removed since it is a form of help for the parents. This way it was easier for a respondent to answer, because it can be that the school helps the parents with the education of their children, but this was not done with a training. All these changes should have resulted in higher consensus between teachers of the same school.

In addition, three statements concerning students reflecting on their own progression were removed. The advisors of OCGH Advies indicated that this was not realistic. Nevertheless, there were still enough statements for the category instructional strategies related to the subcategory *setting objectives and providing feedback* indicated by Marzano (2003).

Another adjustment was related to the structure of the quality scan. The respondents were concerned that their answers would be recognizable by the principal. This was especially related to the category *leadership* since this regards the quality of the work done by the principal. Therefore, in the introduction of the quality scan, the anonymity of the answers became more explicit. Moreover, for the category leadership extra adjustments were made. First, the description of the category was changed. Instead of the term *leadership capacity* the term *leadership style* was used. This made it less obvious that the principal was performing poorly when he or she was receiving a low score. Second, the category was not presented at the end of the quality scan, but in the middle. Now, the respondents could not remove the last page of the quality scan.

Finally, the names of some categories were adjusted. The original translations not always explained the underlying content appropriately. The final names with their original English names are presented in Table 4.1. The sequence of the categories is similar to the sequence represented in the quality scan.

Original English term	Dutch term
Guaranteed and viable curriculum	Haalbaar en gedegen curriculum
Challenging goals and effective feedback	Uitdagende doelen en effectieve feedback
Parent and community involvement	Betrokkenheid van ouders
Safe and orderly environment	Veilige en ordelijke omgeving
Student motivation	Leerling motivatie
Leadership	Leiderschap
Collegiality and professionalism	Professionaliteit en collegialiteit
Teacher motivation	Motivatie
Classroom curriculum design	Lesvoorbereiding
Instructional strategies	Lesuitvoering
Classroom management	Klassenmanagement

Table 4.1: the Dutch terms for the original English categories.

4.7 Summary of design process

Following the previous sections, a lot of changes had to be made to the original questionnaire of Marzano (2003). In Table 4.2, the structure of two statements is presented after each step. This gives an indication on how much the statements have been changed.

Original	A (In my school) a system for early detection of students who are prone to violence and extreme behaviour has been implemented. *(Safe and orderly environment)* B (Teachers in my school) have comprehensive and well-articulated rules and procedures for general classroom behaviour, beginning and ending the period or day, transitions and interruptions, use of materials and equipment, group work, and seatwork. *(Classroom management)*
Step 1 (traditional translation)	A (Op mijn school) een systeem is ingevoerd voor vroeg observatie van leerlingen die de neiging hebben tot geweld en extreem gedrag. B (Leraren op mijn school) hebben allesomvattende en duidelijk verwoorde regels en procedures voor het normale gedrag in de klas, het beginnen en eindigen van een lesuur of dag, overgangen en onderbrekingen, gebruik van materialen en apparatuur, groepswerk en individueel werk.
Step 1 (committee translation)	A (Op mijn school) is een systeem ingevoerd voor de vroege signalering van leerlingen die de neiging hebben tot geweld en extreem gedrag. B (Leraren op mijn school) hebben uitvoerige en duidelijk verwoorde regels en procedures voor het normale gedrag in de klas, het beginnen en eindigen van een lesuur of dag, de overgangen en onderbrekingen, het gebruik van materialen en apparatuur, het groepswerk en zelfstandig werk.
Step 2	A (Op mijn school) is een systeem ingevoerd voor vroegtijdige signalering van leerlingen met een leerprobleem. (*Professionalism and collegiality)* B1 (Leraren op mijn school) hebben heldere regels en procedures voor het normale gedrag in de klas, het gebruik van materialen en apparatuur, het groepswerk en zelfstandig werk. B2 (Leraren op mijn school) hebben heldere regels en procedures voor het beginnen en eindigen van een les of dag en de overgangen en onderbrekingen daarbinnen.
Step 5	A (Op mijn school) wordt het leerlingvolgsysteem gebruikt voor vroegtijdige signalering van leerlingen met een leerprobleem. *(Challenging goals and effective feedback)* B1 Ik hanteer regels en routines voor het gedrag in de klas (bijv. voor het gebruik van materialen en apparatuur, tijdens het groepswerk en tijdens zelfstandig werk). B2 Ik hanteer regels en routines rondom (verschillende onderdelen van) het dagritme.

Table 4.2: Two examples of the adjustments of statements during the design process.

In conclusion, related to the second research question about the number of alterations of Marzano's (2003) questionnaire, it can be argued that a lot of adjustments were necessary. Marzano's questionnaire consisted of 66 statements, while the quality scan was composed of 86 statements (excluding the statements of the general information) after the design process. In total 14 statements were deleted, 28 statements were added, and 6 statements were split up. In addition, 4 statements were reallocated to a different category and 2 categories were removed. This means that 44 percent of the statements used in the quality scan had no direct relation with the original questionnaire of Marzano.

To understand the quality of Marzano's questionnaire, the grounds for the changes will be specified. The differences between the United States and Dutch Primary education, and the statements meant for the Secondary Education resulted in 50 percent of the deleted the statements. The other half was deleted because their content was not clear or because the statement had no real meaning. The new categories and subjects included in the quality scan produced 75 percent of the added statements. The remaining 25 percent of the statements were added, because they specify the content of an original category in more detail. Finally, all six statements were split up, because the statements had a double meaning. Thus, 35 percent of the changes (15 percent of the total statements) were made, because they were inadequate stated by Marzano. The other changes resulted from the inclusion of other categories and subjects, and the differences between the education in both countries. These changes were not a direct effect of bad statements used by Marzano. Nevertheless, with respect to research questions 1 and 2, it can be concluded that without substantial changes Marzano's questionnaire is not applicable in the Dutch Primary Education.

5 Validating the quality scan

In this chapter, the validation of the quality scan will be described. The previous chapter described how the quality scan was created. Here, the analysis of the quality scan will be provided. First, the sample used in this study will be presented. Subsequently, the study's measurements, consisting of the quality scan and several performance measures, will be described. Finally, the statistical methods for the analysis and their results will be presented. Each analysis will be presented in a different section, in which first the specifications of the method are indicated and afterwards the results are provided.

5.1 Sample

The schools that participated in this study were approached in several ways. First, the quality scan was offered as an option to schools in the work region of OCGH Advies with other products. It was part of a larger project of OCGH Advies. This has resulted in three participating schools. Second, the advisors of OCGH Advies received an e-mail with an explanation of the study, its objectives, and the question whether they were executing a project at a school for which implementation of the quality scan could be suitable. One advisor came with a school outside the region of OCGH Advies. After a meeting with the advisor, and me, the principal was enthusiastic and came up with a second school. In addition, two other advisors both knew a school which was willing to participate. Thus, four schools were obtained through advisors of OCGH Advies. Finally, a letter was sent to 139 schools outside the region of OCGH Advies. These schools were in the regions of 's-Hertogenbosch and Tilburg, since these were not too far away from the region of OCGH Advies. This had the advantage that the schools' cultures would not differ too much from the schools in the district of OCGH Advies. After two weeks all schools were called to ask if they had received the letter and if they were interested in participating with the study. This resulted in two schools participating in the study. As a result, the total sample consisted of nine schools. The schools are situated in seven different cities of Noord-Brabant.

In total, 140 individuals, divided over these nine schools, completed the quality scan. The data was analyzed for missing data. Missing data consisted of two types; *ignorable missing data* and *non-ignorable missing data*. The ignorable missing data is caused by the structure of the quality scan. Depending on their function, employees did not have to fill in all the statements. Therefore, for each statement it was indicated what the expected percentage of respondents was. After indicating the expected ignorable missing data with frequency tables in SPSS, the non-ignorable missing data was analyzed. Three respondents were deleted, since their non-ignorable missing data was more than ten percent (Hair et al., 2006). To resolve remaining non-ignorable missing data issues, *mean substitution* was used, which was conducted for 0,6 percent of the data. The mean of the statement of all the respondents of one school was used as the value to be inserted for the missing data[2]. There was no statement that had more than ten percent missing data (maximum was 5,6 percent), so all statements were included for the analyses. As a result, 137 fully completed quality scans were used for the analyses. On average, there were 15,2 respondents per school with a minimum and maximum of respectively 6 and 27. Most of the respondents were female (70,8 percent). The respondents had various functions; 87,6 percent were teachers, 5,8 percent principals, 2,2 percent trainees, and 4,3 percent had different functions[3]. The mean experience of the respondents was 14,1 years (SD = 11,5 years) and if the trainees are excluded this was 14,3 years (SD = 11,5 years).

[2] For comparison of the results of the analyses in SPSS also the dataset without the completion of mean substitution was used.

[3] If a respondent was a teacher with an additional function, he / she is here classified as a teacher.

5.2 Measurements

In the chapters two and four, the categories and the creation of the quality scan were explained respectively. It was argued that the statements should be answered at a five point scale, ranging from (almost) never to (almost) always. In Table 5.1, the number of statements per category is indicated and one or two examples of statements for each category are provided. The statements related to the category *instructional strategies* were divided into four subcategories based on Marzano's work (2003). The category *collegiality and professionalism* was split up into two categories, because a part was answered by all respondents and a part was only meant for teachers.

Category	**Number of statements**	**Example**
A guaranteed and viable curriculum	4	(Op mijn school) wordt de basisleerstof zo geordend dat leerlingen ruimschoots de tijd hebben om deze te leren.
Challenging goals and effective feedback	8	(Op mijn school) worden er specifieke ontwikkel-doelen opgesteld voor de school als geheel.
Parent and community involvement	4	(Op mijn school) is er voor de ouders een lage drempel om contact te leggen met de leerkrachten.
Safe and orderly environment	4	(Op mijn school) wordt door een nette schoolomgeving en duidelijke schoolregels goed gedrag bevorderd.
Student motivation	5	(Op mijn school) nemen leerlingen deel aan leeractiviteiten die betekenisvol zijn.
Leadership	10	(De directeur van deze school) verwoordt een duidelijke visie op de toekomst. (De directeur van deze school) helpt leerkrachten om hun sterke punten te ontwikkelen.
Collegiality and professionalism – General part	5	Ik kom voor mezelf op.
Collegiality and professionalism – Teacher part	3	Ik overleg met andere leerkrachten over de ontwikkeling van mijn leerlingen.
Teacher motivation	4	Ik ben gemotiveerd om mijn werk te doen.
Classroom curriculum design	8	(Bij het plannen van mijn instructieonderdelen heb ik) concrete voorbeelden uitgewerkt die de leerstof verduidelijken.
Instructional strategies – input experience	4	Bij nieuwe leerstof activeer ik de voorkennis van leerlingen door vragen te stellen.
Instructional strategies – regular unit intervals	6	Ik benadruk bij leerlingen het belang van inzet.
Instructional strategies – reviewing, practicing, and applying content	9	Ik geef opdrachten op waarbij leerlingen de leerstof moeten koppelen aan voorgaande leerstof.
Instructional strategies – homework	3	Ik geef individuele leerlingen specifieke feedback over het gemaakte huiswerk.
Classroom management	9	Ik zorg ervoor dat leerlingen vertrouwen krijgen in mijn sturing en leiding.

Table 5.1: Measurements of the quality scan.

Besides the independent measurements, there were also some dependent measurements used during this study. All the dependent variables are at the school level, which will influence the analysis of the data as can be seen in the next section. The first dependent variable was *Cito score*. The Cito test is meant for students of the eighth grade and is used at most of the schools in the Netherlands. The nine schools of this study all participated in the Cito test. The Cito score was gathered with the school results of the last three years. Each school was compared with schools that had an equal population. This population followed from the school reports that each school had from the Cito test. The Cito score was expressed in the difference with the population mean expressed in the number of standard deviations. The equation used for calculating this Cito score is:

$$Cito\ score = \frac{1}{3}\sum_{i=1}^{3}\frac{X_i - \mu_{ip}}{\sigma_{ip}}$$

With: i = 1 to 3 representing 2006, 2007, 2008.
X_i = average score of the school for year i.
μ_{ip} = average score of the population for year i.
σ_{ip} = standard deviation of the population for year i.

The Cito scores of the nine schools had an average value of -0,07 (SD = 0,40) and all Dutch primary schools together had an average value of zero (SD = 1), since it is an standardized normal distribution.

Second, the average level of Secondary Education (*ASE*) the students followed after finishing their Primary Education was used as a dependent variable. The Secondary Education was divided into seven groups. In Appendix A, the different groups are presented. For each group, the percentage of students of the school that went to this level was calculated. The percentage of group one was multiplied with one, the percentage of group two with two, etcetera. These values were summed for all the groups and divided by hundred. This resulted in an average value, which ranged from one to seven. The average of the values of the last two years represented the *ASE* (mean = 4,13; SD = 0,48).

Third, the results of national standard tests of all groups of the Primary Education were used as a dependent variable. These national standard tests were tests of the "leerlingvolgsysteem" (*LVS)*, which are also a product of Cito. The dependent variable was divided into a reading part and a math part. For both an average value and an incremental value was used as dependent variable. The calculation of these variables is in line with the ASE variable. The exact steps are described in Appendix B. The terms of the variables are *LVS average reading* (M = 3,76; SD = 0,23), *LVS delta reading* (M = -0,22; SD = 0,23), *LVS average math* (M = 3,78; SD = 0,18) and *LVS delta math* (M = -0,08; SD = 0,12).

Finally, the factors of the Dutch Education Authority were used as dependent variable. For each school, the last school report of the Dutch Education Authority was requested. As aforementioned, the report consisted of different indicators per factor. The four point scale had a range from *contribute not or barely* (anchor one) to *contribute to a large degree* (anchor four). Since not all indicators were the same for each school, but the schools had a value on almost each factor, the average of the indicators per factor was used as a variable. The factors were already described in section 1.1 (see Table 1.1).

5.3 Results

In this section, the analyses used during this study will be described and the results of these analyses will be presented. For some analyses the full process and its results will be presented. However, for a couple of analyses the total process was quite extensive. Therefore, an example of several process steps of these analyses will be given with all the results.

5.3.1 Reliability Analysis

The reliability of the different categories of the quality scan was calculated with the reliability coefficient Cronbach's Alpha, which assessed the consistency within categories (Hair et al., 2006). The reliability of the categories was both measured for the dataset with and without mean substitution. Since there were no significant differences between both datasets, only the reliability coefficients of the dataset with mean substitution are provided in Table 5.2. A lower limit of 0,60 of the Cronbach's Alpha is commonly used in explorative research, but a value of 0,70 is more desirable (Hair et al., 2006). In general, the reliability of the categories were adequate with values above 0,70 for most categories. Moreover, the Cronbach's Alpha of *collegiality and professionalism (teacher part)* was too low. Nevertheless, the category was included in the correlation matrix and the exploratory factor analysis. It was expected that this category would not form a factor after the factor analysis was conducted, yet it was possible that these statements would form a factor with statements from a different category. Besides, the sample is relatively small; therefore the reliability coefficient will possibly increase with a larger sample.

Category	**N. of cases**	**N. of items**	**Cronbach's Alpha**	**Improved Alpha**	**N. of items deleted**
Guaranteed and viable curriculum (GC)	137	4	0,69		
Challenging goals and effective feedback (GF)	137	8	0,79	0,81	2 (S11, S12)
Parent and community involvement (PI)	137	4	0,63		
Safe and orderly environment (SE)	137	4	0,70		
Student motivation (SM)	137	5	0,70		
Leadership (L)	137	10	0,95		
Collegiality and professionalism (CP)	125	8	0,73		
General part (CP_G)	137	5	0,67		
Teacher part (CP_T)	125	3	0,55	n.p.	
Teacher motivation (TM)	137	4	0,67		
Classroom curriculum design (CC)	124	8	0,78		
Instructional strategies (IS)	49	22	0,87		
Regular unit intervals (IS_RUI)	124	6	0,74		
Input experiences (IS_IE)	124	4	0,75		
Reviewing, practicing, and applying content (IS_RPA)	104	9	0,80	0,83	2 (S78, S79)
Homework (IS_HW)	49	3	0,68	0,76	1 (S82)
Classroom management (CM)	124	9	0,82		

n.p. = not possible

Table 5.2: Reliability analysis of categories.

5.3.2 Exploratory Factor Analysis

An exploratory factor analysis enquires the data and provides information how many factors are necessary to best represent the data (Hair et al., 2006). However, there are some assumptions that should be tested before an EFA can be executed. The first assumption is the degree of normality of the different items. This was tested with the statistical values of the kurtosis and skewness of the items. There were some items that had a statistical value exceeding 1,96 (corresponding to a 0,05 error level). However, the items were still used, because Hair et al. (2006) argued that this is especially problematic with sample sizes below 50.

Second, a Pearson correlation matrix of the different categories of the quality scan was created in Table 5.3, to indicate the relations between different categories. This could be useful for making a choice which categories should be put together in an EFA, since it was not possible to implement all items in one EFA, because of the relatively small sample size.

			A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Group 1	Guaranteed and viable curriculum	A	(0,69)														
	Goals and Feedback	B	,512**	(0,79)													
	Professionalism (general)	C	,338**	,436**	(0,67)												
Group 2	Parent involvement	D	,276**	,288*	,414**	(0,63)											
	Safe and orderly environment	E	,342**	,353**	,439**	,435**	(0,70)										
	Student motivation	F	,467**	,352**	,449**	,473**	,480**	(0,70)									
Group 3	Leadership	G	,252**	,418**	,474**	,309**	,374**	,206*	(0,95)								
	Teacher motivation	H	,175*	,284**	,380**	,308**	,314**	,241**	,330**	(0,67)							
Group 4	Professionalism (teacher)	I	,413**	,391**	,468**	,153	,347**	,330**	,272**	,372**	(0,55)						
	Classroom curriculum design	J	,412**	,465**	,488**	,337**	,432**	,565**	,271**	,409**	,441**	(0,78)					
	Classroom management	K	,363*	,358**	,323**	,212*	,412**	,412**	,181*	,312**	,301**	,503*	(0,82)				
Group 5	Regular unit intervals	L	,255**	,345**	,438**	,210*	,320**	,435**	,286**	,374**	,363**	,559**	,372**	(0,74)			
	Input experience	M	,246**	,292**	,425**	,297**	,381**	,456**	,230*	,254**	,198*	,470**	,400**	,625**	(0,75)		
	Reviewing, practicing & applying	N	,176	,227*	,365**	,311**	,265**	,463**	,334**	,341**	,298**	,517**	,207*	,534**	,429**	(0,80)	
Group 6	Homework	O	,151	,373**	,350*	,353*	,192	,248	,253	,318*	,045	,303*	,285*	,422**	,308*	,482**	(0,68)

Notes: N = varying (49 – 137 respondents)
** Correlation is significant at the 0,01 level (two-tailed)
* Correlation is significant at the 0,05 level (two-tailed)
At the diagonal the Cronbach"s Alpha's are presented

Table 5.3: Pearson correlation matrix of categories.

Hair et al. (2006) specified there should be at least five times as many observations as items analyzed, and a ratio of ten to one would be preferable. Since there were 88 items and only 137 observations, the EFA should be divided in different groups. The correlations between almost all the categories were significant. Therefore, using the correlation matrix, all combinations of categories could be implemented in the same group. The subdivision of the categories was therefore mainly based on my theory-based assumptions regarding which factors suited each other most. The items at the school level were put together as well as possible and the same holds for the items at the teacher level. The categories per group are presented in Table 5.4. Group 6 consisted only of the subcategory *homework*, because only a small part of the respondents had to answer the statements related to this subcategory (N = 49). Thus, the small sample followed from the systematic missing data of the quality scan. If it was combined with another category too much information was lost, due to the listwise deletion of the EFA.

Group	**Categories**	**Number of items**
Group 1	Guaranteed and viable curriculum Challenging goals and effective feedback General part of collegiality and professionalism	17
Group 2	Parent and community involvement Safe and orderly environment Student motivation	13
Group 3	Leadership Teacher motivation	14
Group 4	Teacher part of collegiality and professionalism Classroom curriculum design Classroom management	20
Group 5	Regular unit intervals Input experiences Reviewing, practicing, and applying content (All three subcategories of instructional strategies)	19
Group 6	Homework (subcategory of instructional strategies)	3

Table 5.4: Groups used for the explorative factor analysis.

Finally, the overall correlation within a group was tested, since a minimum level of correlations between the items is necessary for an adequate EFA. These intercorrelations had to be analyzed with the Bartlett test of sphericity and the measure of sampling adequacy (MSA). The Bartlett test of sphericity should be significant and the MSA should exceed 0,50 for an appropriate degree of intercorrelations (Hair et al., 2006). Both measures had a sufficient value for all groups.

Since all the tests of the assumptions yield proper results, the EFA can be executed. Nevertheless, some choices about characteristics of the EFA should be made. First, there are two types of extraction methods; component analysis models and common factor models. The component analysis models are preferred if the theoretical applications are not completely clear (Hair et al., 2006). The quality scan was based on Marzano's questionnaire, which was the result of a meta-analysis. However, the categories of Marzano's questionnaire were not tested in practice. Therefore, theoretical applications were not fixed yet and the principal component analysis in SPSS was chosen as extraction method. Second, a choice between an orthogonal factor rotation and oblique factor rotation should be made. The oblique factor rotation is best applicable if correlations among the factors are expected. Based on the correlation matrix of Table 5.3, this is clearly the case. Therefore, the EFA is carried out using Direct Oblimin in SPSS. These results will be presented. However, the EFA with VARIMAX, an

orthogonal factor rotation, was also executed. Both models were used to indicate the next step of the EFA. It should be noted that for almost all final EFAs the results of both rotation types were the same. During the EFAs, the factor loadings should be at least 0,40 and an item should not have a cross-loading with another factor above 0,40 (Hair et al., 2006). In this report, the exploratory factor analyses of group two will be provided. The other groups followed the same process. In Table 5.5, the first EFA of group 2 is presented. For this EFA all statements of the three groups were included.

Item	**Component**		
	1	2	3
SM27 Feedback on knowledge increase	0,83		
SM28 Meaningful learning activities	0,76		
SM29 Self invented projects	0,50	0,43	
SM31 School wide education	0,48		
PI22 Helping parents by raising their child			
PI20 Low barrier for parents to contact school		0,78	
PI21 Parental involvement in daily activities		0,76	
PI19 Contact teachers with parents		0,67	
SM30 Social involved activities		0,40	
SE25 Consequence violations of school rules			-0,83
SE24 Clear school rules formulated			-0,79
SE23 Promotion of good behaviour			-0,58
SE26 Improvement of self-discipline			-0,49
Eigenvalue	2,97	2,73	2,89

Notes: SM = student motivation, PI = parent and community involvement, SE= safe and orderly environment. Extraction = principal component analysis; rotation = Direct Oblimin.

Table 5.5: Exploratory factor analysis of group 2

From Table 5.5, it followed that structure is not optimal for these items. Item SM29 had a cross-loading with factor two, item PI22 had no loading above 0,40, and item SM30 had a loading on a factor, which was not theoretically related. In addition, for the VARIMAX rotation, item SM30 had a cross-loading with factor one. Therefore, item SM30 was first deleted from the EFA. The EFA without item SM30 was still not optimal. The item PI22 had no loading with the Direct Oblimin rotation and a loading on the wrong factor with the VARIMAX rotation. As a result, item PI22 was deleted next. At the third EFA of group two, only the item SM29 had a cross-loading with factor two. Table 5.6 provides the factor loading matrix without item SM29. The other two matrices are presented in Appendix C.

The factor loading matrix formulated in Table 5.6 had an optimal structure. Hence, all loadings were above 0,40 and there were no cross-loadings present. The items related to factor three had negative loadings, but this does not make a difference for the matrix structure. For the three factors an appropriate name was formulated and the Cronbach's Alphas for these factors were calculated. The results hereof are provided in Table 5.7. The Cronbach's Alpha of the factor *student motivation* was too low. To resolve this issue, item SM29 was added and an analysis of *student motivation* including this item showed that the Cronbach's Alpha was now above the critical value of 0,60. A choice should be made between the reliability and validity of the model. Therefore, for the confirmatory factor analysis both models will be implemented and after that a decision will be made whether to include or delete item SM29 from this factor. The other two items were not used for the confirmatory factor analysis. However, this does not imply that these items should be deleted from the quality scan. Both

items provide specific information about a school. During the discussion, arguments are given on how these individual items can be of value for OCGH Advies.

Item		Component		
		1	2	3
SM27	Feedback on knowledge increase	0,86		
SM28	Meaningful learning activities	0,79		
SM31	School wide education	0,48		
PI21	Parental involvement in daily activities		0,82	
PI20	Low barrier for parents to contact school		0,78	
PI19	Contact teachers with parents		0,70	
SE25	Consequence violations of school rules			-0,85
SE24	Clear school rules formulated			-0,79
SE23	Promotion of good behaviour			-0,51
SE26	Improvement of self-discipline			-0,50
Eigenvalue		2,26	2,13	2,37

Notes: SM = student motivation, PI = parent and community involvement, SE= safe and orderly environment. Extraction = principal component analysis; rotation = Direct Oblimin.

Table 5.6: Final exploratory factor analysis matrix of group 2

Factor name	Included items	N. of cases	Cronbach's Alpha
Student motivation	SM27, SM28, SM31	137	0,55
Student motivation with SM29	SM27, SM28, SM31, SM29	137	0,62
Parent involvement	PI19, PI20, PI21	137	0,63
Safe environment	SE23, SE24, SE25, SE26	137	0,70

Table 5.7: Reliability Analysis of group 2

The process described above of group two was also accomplished for the other groups. Afterwards, the final factor models were tested for the dataset without mean substitution. Hence, in general, this did not provide any differences. Consequently, the factors following from the EFAs could be used as input for the confirmatory factor analyses. The different factors with the number of items per factor and the Cronbach's Alphas are provided in Table 5.8. It was indicated that the factors *student care system* and *personal development* did not reach the critical value for the Cronbach's Alpha. However, this was an explorative study with a limited sample size. It is possible that the reliability will increase if more data are available. Therefore, these factors will be included in the confirmatory factor analysis, but conclusions based on these factors should carefully be handled.

5.3.3 Confirmatory Factor Analysis

With the exploratory factor analysis the items were allowed to load on each factor. This gave a first impression of how the different items are related to each other. The next step was to indicate whether this first impression provided a model having adequate fit. This indication is given by a confirmatory factor analysis (CFA). With a CFA, both the number of factors and on which factor an item will load highly on should be specified before the analysis can be conducted (Hair et al., 2006). In this study, the number of factors and which items should load on which factor for each group were based on the exploratory factor analysis and the theoretical background. Thus, for the CFA it was not the analysis that decided what item loaded on a factor, but the relations were indicated beforehand.

Factor name	# items	N. of cases	Cronbach's Alpha
Learning goals (school)	4	137	0,78
Team professionalism	4	137	0,64
Student care system	2	137	0,59
Education curriculum	3	137	0,69
School development goals	2	137	0,84
Student motivation	3 / 4	137	0,55 / 0,62
Parent involvement	3	137	0,63
Safe environment	4	137	0,70
Leadership	10	137	0,95
Teacher motivation	4	137	0,67
Personal development	2	125	0,51
Pedagogical behaviour	5	124	0,79
Class differentiation	3	124	0,65
Class rules	3	124	0,70
Les planning	3	124	0,66
Practicing content	6	104	0,83
Approach knowledge obtainment	4	104	0,73
Goals and feedback (class)	4	124	0,69
Homework	3	49	0,68

Table 5.8: Factors after exploratory factor analysis

The CFA was accomplished with the software program Lisrel 8.50. The program provides a large number of fit indices for indicating the fit of the implemented model. Five of these indices were used during this study, which represent together all different types of fit measures. First, the Chi-Square is a measure of the differences between the estimated covariance matrix and the actual observed covariance matrix, the smaller the difference, the better the fit. The p-value indicates whether or not this difference is significant; it preferably exceeds 0,05. Furthermore, the degrees of freedom are given, representing the amount of information available to estimate the model parameters. A remark is that with large samples, the Chi-Square is almost always significant, which does not necessarily imply a poor fit of the model. Therefore, a second fit index was used. The Root Mean Square Error of Approximation (RMSEA) is one of the absolute fit measures; it attempts to correct for the tendency of the Chi-Square to reject models with large samples. An adequate fit is found if the RMSEA is smaller than 0,10 and a good fit if it is smaller than 0,05. Third, the Comparative Fit Index (CFI) is an improved version of the incremental fit indicator Normed Fit Index. The CFI evaluates the difference of the Chi-Square for the fitted model and the null model. It is normed so that the values range from zero to one. Values above 0,90 indicate an adequate fit and above 0,95 a good fit. Finally, the Goodness of Fit Index (GFI) and its adjusted version (AGFI) were used, which are both absolute fit indices. The latter takes the model complexity into account. In general, both should exceed 0,90 for a good fit, but the AGFI can be smaller and still indicating a good fit.

As with the EFA, for the CFA one group will be used as an example for the explanation of the process. This group is group one. The first CFA for the group was equal to the final model of the EFA. The Lisrel syntax to create the model is provided in Appendix D and the model itself is presented in Table 5.9. The loadings presented in the table are standardized loadings.

Factor	Item		Loadings	t-value
Fit Indices	Chi-Square = 136,53 (d.f. = 80, p-value = 0,0001); RMSEA = 0,063; CFI = 0,946; GFI = 0,890; AGFI = 0,840			
1 *Education curriculum*				
	GC7	Knowledge of essential content	0,58	6,40
	GC8	Organisation of content	0,80	9,16
	GC10	Realisation effective learning time	0,63	7,09
2 *Student care system*				
	GF11	System providing feedback about students	0,32	1,99
	GF12	System for indication learning problems	1,32	2,33
3 *Learning goals (school)*				
	GF13	Individual schedules with goals	0,72	8,95
	GF14	Individual schedules used for actions	0,78	10,00
	GF15	Group schedules with goals	0,67	8,18
	GF16	Group schedules used for actions	0,67	8,13
4 *School development goals*				
	GF17	Formulation of school development goals	0,84	9,95
	GF18	Development goals used for policy plans	0,87	10,25
5 *Team professionalism*				
	CP_G42	Norms for professionalism	0,75	8,13
	CP_G44	Providing constructive feedback	0,61	6,62
	CP_G45	Be open to receive feedback	0,53	5,64
	CP_G46	Stand up for myself	0,29	2,98

Note: t-value > 1,96 indicates that the factor loading is significant at p<0,05.

Table 5.9: Confirmatory factor analysis group 1 with factors equal to EFA.

The model of Table 5.9 had an adequate fit, concerning the different indices. The RMSEA is below 0,10, the CFI almost 0,95 and the GFI almost 0,90. However, item GF12 had a factor loading above one. This could indicate a problem with the used data (Hair et al., 2006). Item GF11, belonging to the same factor, had the same sign (both positive), which was as expected. According to Hair et al. (2006), it would therefore be hard to conclude whether this was a great problem. However, the reliability coefficient of the factor *student care system* did not exceed the critical value. Therefore, it was chosen to delete both items. The model without *student care system*, given in Appendix E, still had a small loading (0,29) for the item CP_G46, while the t-value was significant (2,93). For the last CFA of this group, the item CP_G46 was deleted. The results hereof are represented in Table 5.10. It can be concluded that this model had an adequate fit looking at the RMSEA and CFI. In addition, all loadings were above 0,50 and had a significant t-value.

After the CFA of this group, the factors used during the next analyses should be chosen. If the goal of this study was purely scientific, the model of Table 5.10 should be used, since the other models had some small problems at a few items. However, this is a first exploratory study and therefore not only statistical issues should be concerned. The fit indices of the three groups are put together in Table 5.11. It followed that the models with and without the item CP_G46 had the almost the same fit. Additionally, looking at the content, the item itself provided some relevant information. Therefore, it was chosen to incorporate this item in the factor *team professionalism.* However, if more data is available in the future, it should again be analyzed what the factor loading of this item is. In case there is a sample of thousand observations and the loading is still this low, the item should be deleted from the factor. Moreover, from a statistical viewpoint, the factor *student care system* should not be used. However, the practical relevance of both items is high in the Dutch Primary Education system.

Therefore, they will be used during further analysis, but results concerning this factor should be interpreted with care.

Factor	**Item**		**Loadings**	**t-value**
Fit Indices	Chi-Square = 85,44 (d.f. = 48, p-value = 0,0007); RMSEA = 0,063; CFI = 0,959; GFI = 0,917; AGFI = 0,866			
1 *Education curriculum*				
	GC7	Knowledge of essential content	0,55	6,09
	GC8	Organisation of content	0,80	9,06
	GC10	Realisation effective learning time	0,65	7,25
2 *Learning goals (school)*				
	GF13	Individual schedules with goals	0,72	8,87
	GF14	Individual schedules used for actions	0,80	10,20
	GF15	Group schedules with goals	0,66	8,05
	GF16	Group schedules used for actions	0,67	8,12
3 *School development goals*				
	GF17	Formulation of school development goals	0,84	9,92
	GF18	Development goals used for policy plans	0,87	10,28
4 *Team professionalism*				
	CP_G42	Norms for professionalism	0,78	8,26
	CP_G44	Providing constructive feedback	0,58	6,21
	CP_G45	Be open to receive feedback	0,50	5,31

Note: t-value > 1,96 indicates that the factor loading is significant at p<0,05.

Table 5.10: Confirmatory factor analysis group 1 without factor "student care system" and CP_G46.

Model	**Chi-Square** (d.f. ; p-value)	**RMSEA**	**CFI**	**GFI**	**AGFI**
Factors of EFA	136,53 (80 ; 0,0001)	0,063	0,946	0,890	0,840
Model without *student care system*	96,90 (59 ; 0,0014)	0,057	0,959	0,912	0,865
Model of Table 5.10	85,44 (48, 0,0007)	0,063	0,959	0,917	0,866

Note: d.f. = degrees of freedom

Table 5.11: Comparison models of group 1.

The process of the CFA described for group one was also executed for the other groups. This resulted in the fact that one item was again included at the factor *activating knowledge*, which was deleted during the EFA. This factor followed from several statements of the subcategory *reviewing, practicing, and applying content*. The model fit for each group is described in Appendix E. The mean, Pearson correlations, and Cronbach's Alphas for the factors are presented in Table 5.12. In addition, the dependent variables were included. These variables were measured on the school level. However, each respondent of a school got the value of the whole school. This meant that there were only a few values for each dependent variable, but that these were duplicated for each school. As a result, the presented Spearman's rho correlations between the factors and performance measures are not totally correct, but they give an indication of whether correlations exist between the variables. It can be concluded that there were, as was the case with the abovementioned categories, many and high correlations between the factors. Besides, there were some correlations between the factors and performance measures. This indicated that it might be possible that there are relations between the factors and performance measures on the school level. Nevertheless, it first should be analyzed whether the factors could be aggregated to the school level. The methods and results of these analyses are provided in the next section.

		Mean	Std.	N	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Learning goals (school)	A	4,05	0,71	137	(0,78)																		
Student care system	B	4,45	0,65	137	0,25**	(0,59)																	
Team professionalism	C	4,00	0,52	137	0,35**	0,25**	(0,64)																
Education curriculum	D	4,11	0,60	137	0,46**	0,21*	0,29**	(0,69)															
School development goals	E	4,12	0,83	137	0,44**	0,18*	0,31**	0,37**	(0,84)														
Student motivation	F	3,83	0,62	137	0,34**	0,22*	0,35**	0,43**	0,19*	(0,62)													
Parent involvement	G	4,50	0,51	137	0,21*	0,20*	0,29**	0,26**	0,14	0,32**	(0,63)												
Safe environment	H	4,28	0,50	137	0,33**	0,06	0,40**	0,34**	0,31**	0,45**	0,31**	(0,70)											
Leadership	I	4,04	0,72	137	0,35**	0,14	0,42**	0,24**	0,40**	0,17*	0,26**	0,37**	(0,95)										
Teacher motivation	J	4,75	0,32	137	0,20*	0,05	0,39**	0,12	0,36**	0,23**	0,22**	0,31**	0,33**	(0,67)									
Personal development	K	4,45	0,57	125	0,25**	0,03	0,42**	0,30**	0,30**	0,21*	0,11	0,31**	0,25**	0,35**	(0,51)								
Pedagogical behaviour	L	4,68	0,37	124	0,31**	0,14	0,32**	0,27**	0,32**	0,35**	0,11	0,27**	0,23*	0,27**	0,23*	(0,79)							
Class differentiation	M	4,12	0,58	124	0,30**	0,13	0,32**	0,26**	0,19*	0,44**	0,19*	0,28**	0,23*	0,29**	0,29**	0,35**	(0,65)						
Class rules	N	4,65	0,43	124	0,24**	0,12	0,23*	0,32**	0,25**	0,37**	-0,04	0,42**	0,04	0,26**	0,21**	0,51**	0,26**	(0,70)					
Les planning	O	4,17	0,59	124	0,43**	0,19*	0,36**	0,30**	0,31**	0,43**	0,21*	0,36**	0,23**	0,36**	0,20**	0,40**	0,43**	0,28**	(0,67)				
Practicing content	P	2,99	0,78	104	0,13	0,06	0,21*	0,03	0,13	0,32**	0,29**	0,15	0,27**	0,25*	0,06	0,09	0,30**	-0,03	0,29**	(0,83)			
Approach knowledge obt.	Q	4,16	0,52	124	0,34**	0,20*	0,45**	0,26**	0,30**	0,52**	0,34**	0,47**	0,27**	0,34**	0,24**	0,46**	0,46**	0,38**	0,54**	0,43**	(0,75)		
Goals and feedback (class)	R	4,15	0,53	124	0,30**	0,23*	0,47**	0,22*	0,18*	0,38**	0,10	0,30**	0,27**	0,37**	0,24**	0,33**	0,49**	0,27**	0,41**	0,46**	0,56**	(0,69)	
Homework	S	3,19	0,96	49	0,26	0,22	0,34*	0,14	0,33*	0,19	0,35*	0,19	0,25	0,32*	-0,01	0,34*	0,24	0,22	0,22	0,44**	0,40**	0,43**	(0,68)
Cito score		-0,07	0,40	9	-0,13	0,00	-0,18*	0,05	-0,35**	0,20*	0,22*	-0,02	-0,09	-0,08	-0,02	-0,01	0,02	-0,14	0,04	0,04	-0,11	-0,09	-0,03
ASE		4,13	0,48	9	-0,08	0,04	-0,13	0,15	-0,24**	0,30**	0,26**	0,06	0,01	-0,04	-0,08	-0,03	0,02	-0,17	0,03	0,17	-0,09	-0,05	0,02
LVS delta reading		-0,22	0,23	6	0,12	0,11	0,17	0,07	0,00	0,09	0,06	-0,08	-0,18	-0,04	0,19	0,00	0,06	0,17	-0,05	-0,05	-0,04	0,27	0,18
LVS delta math		-0,08	0,12	5	0,20	-0,02	0,16	0,23*	0,22*	0,16	-0,10	0,20	-0,24*	-0,06	0,11	-0,06	-0,02	0,18	-0,06	-0,06	-0,01	0,10	0,13
LVS average reading		3,76	0,12	6	0,18	0,07	0,17	0,02	0,23*	-0,08	-0,09	0,01	-0,29**	-0,06	0,01	0,11	-0,02	0,45**	0,06	-0,13	0,10	0,28**	0,32
LVS average math		3,78	0,18	6	0,02	0,09	0,11	0,03	0,06	0,05	0,02	-0,06	-0,15	0,02	0,08	0,17	0,13	0,30	0,25*	0,08	0,10	0,33**	0,22

Notes: Std. = standard deviation, ASE = average level of Secondary Education.
** Correlation is significant at the 0,01 level (two-tailed)
* Correlation is significant at the 0,05 level (two-tailed)
At the diagonal the Cronbach"s Alpha's are presented
Above dotted line Pearson correlations and below Spearman's rho correlations.
Correlations with performance measures (below dotted line) are indications!

Table 5.12: Correlation matrix of factors and performance measures.

5.3.4 Aggregation

The factors created during both types of factor analyses were all measured on the individual level. However, the individuals did not provide completely independent answers, since several individuals belong to the same school. In this section, it will be analyzed whether the individual answers could be combined to a score for the school as a whole using an aggregation analysis.

The first method used to conduct the aggregation was the within-group interrater reliability (r_{wg}) of James, Demaree and Wolf (1984). With this method, the variance of each group (i.e. school) will be compared with the expected variance of a specific distribution. In general, the expected variance of a Uniform Distribution (EU) is used for comparison (James et al., 1984). For a five point scale, the expected variance is 2,00. However, the distribution of the factors in this study was negatively skewed for almost all factors. This could have been caused by social desirability or positive leniency of the answers. In this case, the Uniform Distribution will give a too optimistic result. Therefore, the expected variance of a small negatively skewed distribution (ESS) was used in this study as well. The expected variance of this distribution is 1,34. The median r_{wg} of the nine schools for each factor in comparison with both distributions is presented in Table 5.13[4]. The within-group interrater reliability of almost all factors had a value above 0,70 for both distributions, which is an appropriate value (Klein et al., 2000). Only the factor *homework* had a value lower than 0,60. This was probably caused by the small number of respondents per school. In addition, this factor had the less skewed distribution, so the value of r_{wg} compared with the expected variance of the Uniform Distribution was most appropriate. This had a median of 0,56, which is a little low, but can increase with a larger sample and therefore this factor was aggregated to the school level.

Factor	**r_{wg} (EU)**	**r_{wg} (ESS)**	$\tilde{n}$	**F-value**	**Sig.**	**ICC (1)**
Learning goals (school)	0,85	0,72	14,83	1,34	0,229	0,02
Team professionalism	0,90	0,83	14,83	1,92	0,062	0,06
Student care system	0,84	0,73	14,83	1,07	0,388	0,00
Education curriculum	0,91	0,84	14,83	1,12	0,356	0,01
School development goals	0,83	0,72	14,83	4,82	0,000**	0,20
Student motivation	0,90	0,83	14,83	4,88	0,000**	0,21
Parent involvement	0,93	0,88	14,83	3,70	0,001**	0,15
Safe environment	0,94	0,90	14,83	2,04	0,046*	0,07
Leadership	0,96	0,93	14,83	5,40	0,000**	0,23
Teacher motivation	0,97	0,96	14,83	0,56	0,810	-0,03
Personal development	0,91	0,86	13,40	1,87	0,071	0,06
Pedagogical behaviour	0,98	0,96	13,40	1,22	0,295	0,02
Class differentiation	0,88	0,79	13,40	1,27	0,265	0,02
Class rules	0,96	0,94	13,40	3,10	0,003**	0,14
Les planning	0,91	0,85	13,40	1,78	0,089	0,05
Practicing content	0,86	0,67	11,28	2,03	0,051	0,08
Approach knowledge obtainment	0,94	0,89	11,28	0,76	0,642	-0,02
Goals and feedback (class)	0,93	0,88	13,40	2,15	0,037*	0,08
Homework	0,56	0,00	5,20	0,74	0,655	-0,05

Note: ** ANOVA test is significant at the 0,01 level
* ANOVA test is significant at the 0,05 level

Table 5.13: Aggregation results for r_{wg} and ICC(1) methods.

[4] The dataset without mean substitution provided the same results. There were no larger differences than 0,02 in comparison with the dataset with mean substitution.

The within-group interrater reliability analysis does not use the between-group variance of the factors, only the within-group variance is used in this method. Therefore, the interclass correlation coefficient (ICC(1)) was also calculated (Bliese, 2000). The advantage of the ICC(1) is that it indicates whether individuals of one group provide more cohesive answers compared to the whole population. Thus, all the data available for a factor were used at once, so conclusions could be made concerning the differences between groups. Besides, in an exploratory study the appropriate level to aggregate the data is not exactly known. Since the absence of between-group variability indicates that the expected group level does not exist, this can provide useful information concerning the appropriate level (Chan, 1998). With this measurement the population variance *between* groups will be divided by the total variance. A large ICC(1) indicates that most of the total variance is caused by the between-group variance. Hence, the respondents within a group provide consistent answers, while respondents of different groups give diverse answers (Bliese, 2000). The ICC(1) was established with the results of the ANOVA tests, using the following equation:

$$ICC(1) = \frac{MSB - MSW}{MSB + \left[(\tilde{n} - 1) \times MSW\right]}$$

With: MSB = Mean square between groups
MSW = Mean square within groups
$\tilde{n}$ = average number of respondents per school.

Since the sample size was different for each school, the average sample size was calculated with the equations of Snijders and Bosker (1999). The results of these tests are also included in Table 5.13[5]. The results of the ICC(1) were less desirable than the results of the r_{wg}. There were seven factors that had significant differences between the schools. Hence, according to the ICC(1), the other factors could not be aggregated to the school level. The difference between the results of both methods could have been caused by the homogeneity of the schools. The schools have comparable populations and are located in the same district. Therefore, there are few differences between the schools and it is likely that the answers between the schools are similar. This resulted in a low ICC(1), but had no effect on the r_{wg}. As a result, the outcomes of the r_{wg} were used and it was assumed that all factors could be aggregated to the school level. However, if data of more diverse schools, for instance very good and bad schools, is available, the ICC(1) should be executed again. With the diverse dataset the results of a factor should be significantly different between the schools. If this is not the case, it can be concluded that the factors cannot be aggregated to the school level. In addition, in Table 5.13 some negative values of the ICC(1) can be seen. This meant that the within-group variability was larger than the between-group variability. These values could be regarded as zero, since the negative values had no meaning.

5.3.5 Relations with performance measures

The analyses of the previous subsection showed that all the factors could be aggregated to the school level. The performance measures in this study were also at the school level. Therefore, in this subsection, the relation between the factors and the performance measures will be analyzed. The strongest conclusions generally follow from a multiple regression analysis. With this analysis, the prediction of the independent variables on the dependent variable can be established (Hair et al., 2006). However, the sample size in this study was too small (N=9 at the school level). Therefore, a different analysis was used for indicating a relation.

Since only a sample of nine cases was available, the results should be handled with care. The nonparametric method Kendall's tau b was used to indicate the relations between the factors and dependent variables. The Kendall's tau b correlates ranks between two ordered variables (Cooper & Schindler, 2003). The performance measures Cito score, average level of Secondary Education

[5] The results of the ICC for the dataset without mean substitution provided one difference with the dataset with mean substitution. The factor *safe environment* was only just insignificant for that dataset.

(ASE), and the four measures of the LVS system were used as dependent variables. The results of the Kendall's tau b test are presented in Table 5.14.

Factor	**Cito Score**	**ASE**	**LVS average reading**	**LVS delta reading**	**LVS average math**	**LVS delta math**
Number of schools	*9*	*9*	*6*	*6*	*6*	*5*
Learning goals (school)	-0,05	0,02	0,75**	0,37*	0,45**	0,42*
Team professionalism	*-0,43***	-0,21	0,73**	0,27	0,55**	0,37
Student care system	-0,16	0,14	0,44*	0,75**	0,46**	0,23
Education curriculum	0,09	0,33**	0,59**	0,21	0,57**	0,34
School development goals	*-0,60***	*-0,32***	0,38*	-0,32	0,11	0,81**
Student motivation	0,33**	0,60**	-0,18	0,20	0,08	0,89**
Parent involvement	-0,11	0,19	0,09	0,01	0,76**	-0,12
Safe environment	0,08	0,11	0,17	-0,21	0,14	0,89**
Leadership	*-0,39***	*-0,45***	*-0,88***	*-0,42**	*-0,49***	-0,23
Teacher motivation	*-0,32***	-0,02	-0,08	0,31	0,60**	-0,06
Personal development	0,03	*-0,25**	0,61**	0,52**	0,67**	0,37
Pedagogical behaviour	0,22	0,65**	0,77**	0,07	0,51**	0,26
Class differentiation	-0,10	0,12	0,24	0,07	0,91**	-0,06
Class rules	0,30**	0,38**	0,77**	0,07	0,51**	0,26
Les planning	0,05	0,27*	0,41*	0,01	0,84**	0,01
Practicing content	-0,12	0,17	*-0,39**	-0,08	0,27	*-0,61***
Approach knowledge obtainment	-0,01	0,34**	0,75**	0,05	0,49**	0,28
Goals and feedback (class)	0,04	0,39**	0,35*	0,65**	0,46**	0,37
Homework	0,12	0,56**	0,68**	0,06	0,42*	0,34

Note: ** Correlation is significant at the 0,01 level (two-tailed)
* Correlation is significant at the 0,05 level (two-tailed)

Table 5.14: Kendall's tau b correlation matrix of factors and performance measures.

The Kendall's tau b correlation matrix specified several significant correlations between the factors and performance measures. However, in contrast with the correlations between the factors, not all significant correlations were positive. There were also twelve negative correlations (i.e. 21 percent) divided over six factors and all performance measures. Based on the literature study, only positive correlations were expected, since for each factor several studies have shown positive effects on the performance measures. The performance measures *LVS average math* and *LVS average reading* provided the best results. Both measures had positive correlations with many factors, and only one and two negative correlations respectively. Especially for LVS average math the relations were strong, since the correlations were almost all significant at the 0,01 level. The *Cito score* had the worst relation with the factors, since five of the six significant correlations were negative. This indicated that performing good on the different factors should result in a poor average Cito score. Moreover, the ASE also showed three negative correlations, yet it also showed eighth positive correlations. The factors *class rules* and g*oals and feedback (class)* provided the best relations with the performance measures. They were significant positively correlated with four performance measures. The remaining two were positive, but not significant correlated. In contrast, *leadership* presented the worst result with five significant negative correlations. This suggests that good leadership of the principal would cause negative student performance. Additionally, *practicing content* also provided only significant negative correlations, while *team professionalism*, *school development goals*, *teacher motivation*, and *personal development* had both negative and positive correlations. The other fourteen factors had only significant positive correlations, although *parent*

involvement, *safe environment*, and *class differentiation* provided a significant correlation with just one of the six performance measures.

The last performance measures were the factors of the Dutch Education Authority. The values of each factor of the Dutch Education Authority were calculated by averaging the values of the indicators for each factor. This resulted in a value for the twelve factors of the Dutch Education Authority for each school. However, after analyzing these values, it followed that there was barely any variance between the values of the schools per factor. The frequency tables showed only for the factor *quality management* six different values, while these values were already averaged using the values of the indicators. Hence, the standard deviation of this factor was still low. The other factors had at most three different values and two factors had even just one value. In addition, all schools scored a value around three for each indicator, which implied a sufficient rate at the factor. With this lack of variance, it was not possible to execute a Kendall's tau b correlation between the factors of the quality scan and the factors of the Dutch Education Authority. Nevertheless, the factors of the Dutch Education Authority were not useless, since it was an additional indication for the homogeneity of the schools in the sample. This fact will be used in the discussion section to explain the results regarding the factors of the quality scan.

In this section, the results of the analyses were presented. This was mainly done from a scientific research perspective. Therefore, in the next chapter, the implications for OCGH Advies will be explained. Subsequently, in Chapter 7, the results will be discussed and the research questions will be answered by means of the results of these analyses.

6 Implementation

In the previous chapter, the quality scan has been analyzed. It was established which statements belong to the same factor. However, for OCGH Advies this will not be the end of the process. Their goals are identifying opportunities for improvement, and to present a clear feedback report to the principal or school team. Therefore, in this chapter, the implementation of the quality scan will be explained. First, it will be indicated how the quality scan should be completed, how the feedback report should be created, and how the feedback should be given to the school. Finally, some links with several products of OCGH Advies will be indicated.

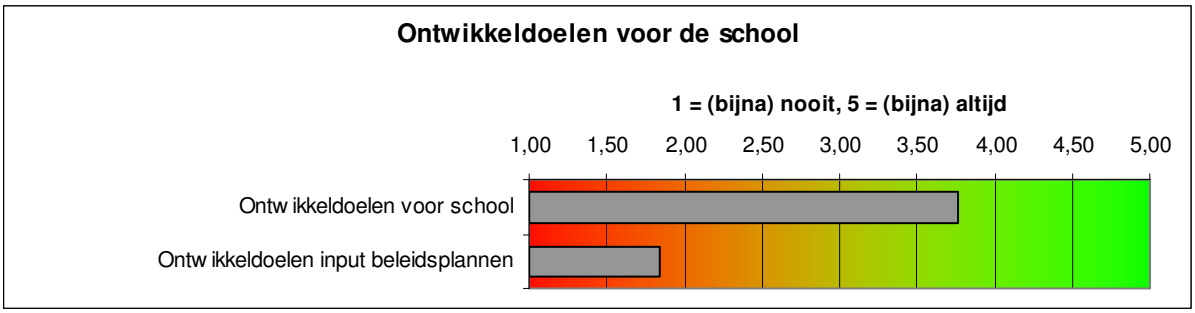
6.1 Use of the quality scan

During the analysis of the quality scan, some statements were deleted from the analysis. However, these statements were not automatically deleted from the quality scan for OCGH Advies. During a meeting with an advisor, it was indicated where the statements that were not included at a factor should fit best. This allocation was based on the content of the statements. Finally, one statement was deleted, since it was not statistically related to a factor and the content had no additional value for the quality scan. The other statements were allocated to a different factor. Hereafter, the structure of the quality scan was changed. Each factor became a category of the quality scan with a separate headline. Thus, the quality scan now consists of 20 categories. Nevertheless, an advisor can choose to delete some categories from the quality scan with the distribution of the quality scan at a school in case the goal of the project is more specific. For example, if a school would like to know how teachers are behaving in the class, the eighth factors at the teacher level will be sufficient.

After indicating which categories should be included by the advisor, the way of collecting the data should be chosen. It is possible to provide an envelope with a copy of the quality scan to each respondent, so their answers will be more anonymous. Another option is that the school team completes the quality scan during a team meeting or that they will get a week to complete the quality scan. With the first option, there is a greater chance that everybody completes the quality scan, while with the second option, they can think longer about each answer at a statement. The advisor should make the choice on how the data should be collected. Therefore, different versions belonging to the different options were provided to OCGH Advies. I would recommend the following option; the school team gets a week to complete the quality scan and they can put their copy in an envelope. Using this process, the respondents can be honest even when their answers are negative towards the school or principal.

After the completion of the quality scan, the data should be reformulated to usable information. An Excel file was developed for this process. Someone of OCGH Advies can enter the answers of all respondents into the Excel file. In Excel, the report will be created. The content of the feedback report is standardized. It starts with a title page and the content of the report is described. In addition, with the input of the data, some graphs will automatically be created in Excel. The first graphs present per category the mean of each statement, indicated by an abbreviation of the statement. Thus, the results of the separate statements will be used, since these will provide unique information for a school. In addition, for each statement the dispersal of the answers will be presented. This is done by providing the percentages of the answers of the five scale points in a table together with the graph of each category. The information per category is completed using a box wherein questions can be inserted by the advisor. These questions should stimulate a discussion in the school team and should mainly focus on the statements with a low score. The statements having a high score can be handled during a

meeting where the feedback report will be discussed. An example of the feedback for a category is presented in Figure 6.1. At the end of the feedback report, the mean of each factor is provided as a summary. Besides, some general points of interest can be formulated. This report can be printed and sent to the school. It is not yet possible to create a benchmark for the different statements or factors, since there are only data from nine schools available. Moreover, it is not necessarily beneficial to include a benchmark, because based on a benchmark a school can conclude that some statements are scoring above average and therefore no improvements are necessary. Thus, I would not recommend to include a benchmark, but if more schools have completed the quality scan, this is possible.



	Ontwikkeldoelen voor de school				
	Percentages antwoorden				
	1	2	3	4	5
Ontwikkeldoelen voor school			38%	46%	15%
Ontwikkeldoelen input beleidsplannen	38%	46%	8%	8%	

Vragen:
- Hoe zouden wij de gestelde ontwikkeldoelen om kunnen zitten in beleidspunten?

Figure 6.1: Example of feedback per category.

After finishing the report, the results should be discussed with the principal or school team. For the interpretation of the results of the quality scan, it is important to keep in mind that the report is the result of the opinion of the school team. Their opinion can be biased, because they evaluate themselves. Therefore, the advisor of OCGH Advies should plan a feedback meeting with the principal or school team. During this meeting, the advisor can ask specific questions to discover extra information concerning the different factors. Hence, the advisor should focus both on the factors that are scoring low and the factors that are scoring high, since for these factors the interpretation of members of the school team can be different. The benefit of the quality scan is that during this meeting the advisor has some starting-points. Besides, if a factor scores low, it is harder for the school team to argue that this is not the case, because it resulted from their opinion in the first place. Thus, they should come with good arguments to indicate that this is not the case. In conclusion, the meeting should indicate the most important points of improvement. For OCGH Advies, this should lead to new projects at the school.

6.2 Link with products

After indicating which factors or statements should be improved at a school, the advisor should know which products are available at OCGH Advies to accomplish this. Therefore, it is useful to indicate the link between the quality scan and products of OCGH Advies. The choice of OCGH Advies was to use the Dutch version of the EFQM model for specifying these linkages.

First, the factors of the quality scan should be classified into different criteria of the EFQM model. Most of the factors belong to the criteria *processes* of the EFQM model. This is logical, since the goal is to identify the quality of the processes at a school. Nevertheless, some factors are related to different criteria. First, the factor *leadership* can be categorized in the criterion *leadership*. Second, both *team professionalism* and *personal development* are classified in the criterion *people management*, since both factors should increase the competence of the team or its members. Third, the factors *student care system*, *education curriculum*, *school development goals*, and *safe environment* are allocated in the criterion *policy and strategy*. All four factors should be implemented at the school level. Ideally, these factors are described in the policy of the school. Fourth, *teacher motivation* is categorized in *people satisfaction*. Although satisfaction and motivation are not similar, both terms are closely related. Therefore, it is decided to link the factor and the criterion. Finally, *parent involvement* and *student motivation* are divided into the criterion *customer satisfaction*, since both the students and parents can be seen as the customers of a school. The classification of the factors into the different criteria is also presented in Figure 6.2.

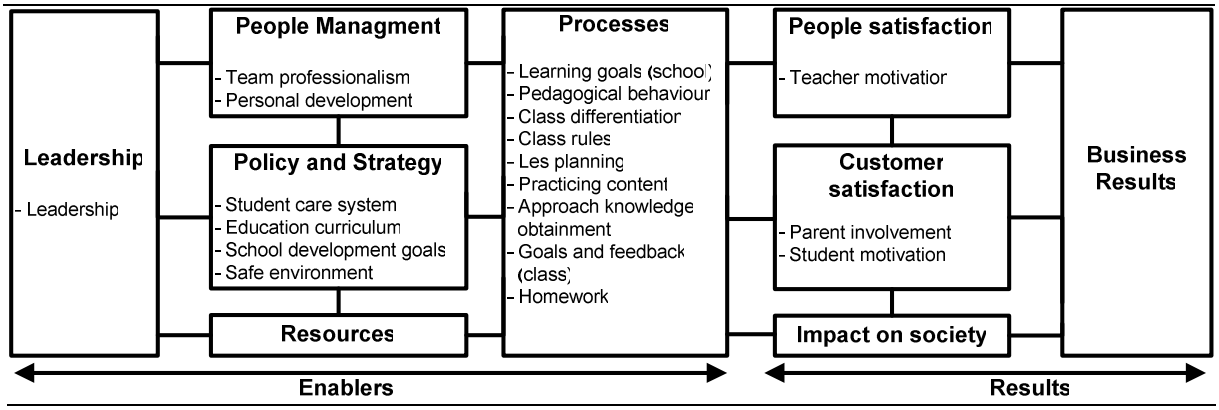


Figure 6.2: EFQM-model with factors of the quality scan.

OCGH Advies has a large range of product available to improve the different factors. Here some examples of products will be presented. First, for the factors *leadership* and *personal development* the products related to *personal coaching* of OCGH Advies will be useful. For example, the development of a personal developments plan ("persoonlijk ontwikkelingsplan") can be beneficial for the principal or team members. Second, for the factors related to the different types of goal setting (*class learning goals, school development goals* and *setting les goals* e.g.), the products "Passend Onderwijs" and "éénzorgroute" are valuable products. Both products focus on goal oriented teaching; which will probably improve the different factors. Subsequently, OCGH Advies provides trainings related to classroom management and contacts with parents, which are related to *class control* and *parent involvement* respectively. Finally, for the factors at the teacher level, the product "taakspel" will be very effective. Working with this product, the teacher is actively working with the students on his / her classroom management and instruction strategies. Here, the focus during the "taakspel" can be put on the different factors at the teacher level of the quality scan.

In summary, it can be concluded that the quality scan can be used by OCGH Advies. It will provide the advisors of OCGH Advies with relevant information of a school. In addition, after the identification of a school's problems, OCGH Advies can use appropriate products to solve the problems of the schools. Furthermore, a business report is written for OCGH Advies. In this report, the use of the Excel file is described in more detail and for each factor more relevant products of OCGH Advies are provided. In the next chapter, discussion, the results of the analysis will be interpreted and possible future research directions will be presented.

7 Discussion

In the previous chapters, the project process and its results have been described. First, the project objective was formulated. Subsequently, relevant literature and research questions were presented. Thereafter, the design and validation processes were described with their results. However, the interpretation of these process steps was not completely provided. Therefore, in this chapter, the implications of the different analyses will be discussed. Different types of validity and how these were achieved for the quality scan will be indicated. Next, the answers to the different research questions will be provided. Thereafter, the research limitations of this study and possible future research directions will be presented. Finally, a conclusion concerning this study and the quality scan will be formulated.

7.1 Validity

The validity of a model is measured using different types of validity. Scandura and Williams' (2003) model consisting of the distinction between four types of validity will be used to determine the validity of the quality scan. First, *statistical conclusion validity* refers to the ability to draw conclusions based on statistical evidence that is dependent on the sample size and the number of dependent variables (Scandura & Williams, 2000). Second, *construct validity* determines to which extend the variables have been adequately defined and measured by methods (Hair et al., 2006). Third, *internal validity* concerns causality, refers to the certainty of a cause-and-effect relationship between the variables in the study (Scandura & Williams, 2000). Finally, *external validity* is the ability to generalize the results of other schools, respondents, and time (Cooper & Schindler, 2003). Each type of validity will be discussed in this section.

7.1.1 Statistical conclusion validity

Statistical conclusion validity refers to the strength of the conclusions that can be drawn from the analysis. This type of validity is strongly connected with internal validity. However, the difference being that internal validity determines whether there are external factors influencing the relationships, while statistical conclusion validity indicates if the analysis' procedure is appropriate to accept the results. As aforementioned, the sample size was relatively small as there were only nine schools and 137 respondents. Nevertheless, the 88 statements were divided in several groups that theoretically were strongest related. The strength of the theoretical relationship was based on several literature sources. Furthermore a clear distinction was made between statements at the school and teacher level. Therefore, there were fewer items per group, which increased the power of the statistical test. Moreover, a clear procedure was followed to analyse the data and the assumptions that were necessary to test the different analyses. In addition, several dependent variables were used. A factor with a positive relation with a number of these variables would increase the statistical conclusion validity. The relations between the factors and performance measures will be described in section 7.1.3, *the internal validity*, but it can be concluded that the statistical conclusion validity of the study was adequate.

7.1.2 Construct validity

Construct validity determines both whether the items were well defined and whether the methods measured the construct of the factors adequately (Scandura & Williams, 2000). The first part is handled by *face validity*, which means that every statement's content or meaning should be understood (Hair et al., 2006). In this study, the importance of the different categories was validated by consulting additional literature. Subsequently, during the design of the quality scan, the statements were handled and discussed with a professor and an advisor of OCGH Advies. Terms of the

statements were chosen with care, so the chance of dissimilar interpretations was reduced. The discussions with different stakeholders and the consulting of a book of the teacher education program improved the understanding of the content of each statement. Thus, it can be concluded that the face validity for this study was good.

The second part of construct validity, concerning the measurement of the factors was executed with different analyses, like the reliability analysis, factor analysis, and aggregation analysis. In general, these analyses provided appropriate results. Nevertheless, some results were not as expected or did not met all requirements of the analyses. These will be discussed here.

The confirmatory factor analysis of each group established an adequate to good fit. The variance extracted for each factor per group were for almost all factors larger than the square of the correlation estimate between this factor and another factor of the group, which specified that the factors are truly distinct from each other (Hair et al, 2006). Although the adequate model fit, the Cronbach's Alphas of the factors *student care system* ($\alpha = 0{,}59$) and *personal development* ($\alpha = 0{,}51$) were below the critical value of 0,60. A possible explanation for the low reliability is that the factors consisted of only two statements, while a factor in general consists of at least four items (Hair et al., 2006). Nonetheless, strictly argued, these two factors cannot be seen as factors, since their reliability is too low. At the other hand, the conducted study was a first exploratory study with a limited sample size, and the practical relevance of the factors was established during conversations with advisors of OCGH Advies. Besides, with a larger sample the Cronbach's Alpha generally increases, so both factors are still included in the quality scan. When more data are available, the reliability of both factors should once more be determined. If these reliabilities are still too low when re-evaluated, they should be deleted from the quality scan.

The interclass correlation coefficient (ICC(1)) also showed some remarkable results. The quality scan included two types of statements. The statements at the school level were stated using a *group-referent* (e.g. in my school students receive feedback on their knowledge growth) and the statements at the teacher level were stated in the first person, which is according the *individual-referent* (e.g. I emphasise the importance of effort to students) (Klein et al., 2001). Five out of nine factors consisting of statements with the *group-referent* had a significant ANOVA test and only two of the ten ANOVA tests for the factors with the *individual-referent* were significant. According to the composition models of Chan (1998), the group-referent belongs to the *referent-shift consensus models*. The statements answered by the individuals were shifted to the group level. Thus, it did not ask how the individual (i.e. lower-level) executed a specific action, but how the action was executed by the school (i.e. higher-level). In contrast, the individual-referent is an example of the *direct consensus models* (Chan, 1998). The assumption on these models is that the construct at the lower-level (i.e. individual level) is equal to another form of the construct at the higher-level (i.e. school level). Thus, using the group-referent, the items and factors were stated at the school level, while using the individual-referent, the items and factors were stated at the teacher level. Furthermore, it was assumed that the relations between the items and factors would be the same at school level (Chan, 1998). This distinction can explain the differences in the results of the ICC(1) for the factors at the school level and teacher level. The few number of significant ANOVA tests with the individual-referent could implicate that the factors did not have the same characteristics at the school level as they did at the teacher level. In addition, the non-significance of the ANOVA tests can be caused by the decrease of the between-group variability or by the increase of the within-group variability. The average between-group variability was 1,24 and 0,48 for the group-referent factors and individual-referent factors respectively, and the average within-group variability was 0,36 and 0,35 for the group-referent and

individual-referent respectively. Thus, the lack of a significant ANOVA test for the individual-referent factors was caused by the decrease in between-group variability. Klein et al. (2001) revealed the same conclusions, since they proved that the between-group variability decreased with an individual-referent. They did however not calculate the ICC(1). Gamero, Romá and Peiró' (2008) study showed that the ICC(1)s were higher for the factors measured with the referent-shift consensus model than for factors measured with a direct consensus model. Thus, the remark made by Chan (1998) that the aggregation of factors to a higher level with direct consensus models should be handled with care is proven in this study. A larger study is still necessary to determine if the individual-referent factors can or cannot be aggregated to the school level. Hence, until this is tested with a larger sample, it is assumed that the individual-referent factors can be aggregated based on the within-group interrater reliability (r_{wg}).

7.1.3 Internal validity

The third facet of validity is the internal validity, which is concerned with the causality between variables (Scandura & Williams, 2000). The study was conducted with a field survey. Therefore, the context cannot be controlled, which reduces the certainty of the causality between the factors and performance measures. Besides, the cause-and-effect relationship cannot be concluded due to the cross-sectional design of the study. The data for the performance measures were gathered in a period before the completion of the quality at the schools. Thus, the best conclusion that can be drawn from the factors and performance measures using this study is that a conditional relationship, but not a casual relationship exists. Nevertheless, the correlations matrix of the factors and the correlation matrix between the factors and the performance provide some evidence for internal validity.

First, the Pearson correlation matrix determined that there were many significant correlations between the factors, which were all positive. Since it was argued that all factors had a positive effect on the student's performance, it was expected that the correlations were positive (Marzano, 2003). Besides, several studies proved that different factors were strongly related. For example, Pressley et al. (2007) showed a strong relation between *leadership* and the different factors of the category *instructional strategies*. The correlation matrix indicated that three factors of the category instructional strategies had a significant correlation with leadership, which is in line with Pressley et al.'s (2007) study. A second example is the study of Reeve and Jang (2007). They identified a positive relation between the student-teacher relation and student motivation. The student-teacher relation is similar to the factor *class control*, which had a significant correlation with the factor *student motivation*. These significant positive correlations enhance the possibility of causality between factors and for that reason enhance the internal validity (Hair et al., 2006).

Second, the Kendall's tau b correlation matrix between the factors and several performance measures could be the first indication that causal relations exist between both. However, in contrast with the correlations between the factors, not all significant correlations were positive. Based on the literature study, only positive correlations were expected, since for each factor several studies have showed positive effects on the performance measures. Most remarkable was that the factor *leadership* was negatively correlated with almost all performance measures. A reason for this can be that a school enlists a new and good principal when the performance of the students is disappointing. The performance measures however did not yet quantify this effect. Another possible explanation is that a principal had introduced changes which were paying off, but that these changes were not appreciated by the teachers. The other negative correlations could all be caused by the time dimension. Since the school were implementing improvements, which had an effect on several factors, because of the disappointing results that were used as performance measures in this study. It followed that poor

results of the performance measures were combined with good scores of some factors. An indication hereof is that most negative correlations were established at the school level. At this level the changes can be implemented the quickest, because these factors deal with rules, while at the teacher level, the factors are regarded behavioural factors, which are more difficult to change.

7.1.4 External validity

The last facet of validity is the external validity of a study. The external validity refers to generalizing the results across time, settings (i.e. schools), and individuals (Scandura & Williams, 2000). A survey study enhances the external validity, since several schools with different settings participated. Nevertheless, there were some indications that the sample consisted of homogeneous schools. The first signal came from the population of the schools. The Cito institute uses seven different types of school groups to represent a school population. This distinction is measured by indicating the social-cultural background (i.e. sociaal-culturele achtergrond) of the students (Terugblik en resultaten, 2008). Five schools of this study are classified as school group two, three as school group three, and one as school group one. Hence, the population of the schools is quite homogeneous. Still, the performance of the nine schools could vary greatly. However, the school reports of the Dutch Education Authority for the schools were very similar. All schools were scoring a sufficient at all the factors of the standardized evaluation tool. In addition, the *Cito scores* of the schools in the sample had a standard deviation of 0,40, while the standard deviation for the whole population is one. This similar performance at the different aspects of the schools was an additional indication of the homogeneity of the schools. Hence, this homogeneity of the schools decreases the *external validity*, since the participated schools are not representative for all the schools in the Netherlands. Nevertheless, it can be argued that the results formulated in this study are representative for schools which consist for at least 50 percent of students coming from a good social-cultural background (i.e. students with a "*leerlinggewicht*" 1,00 (Terugblik en resultaten, 2008)).

In summary, the different facets of validity of the quality scan are demonstrated. Although not for each factor or statement all the criteria were met, it can be concluded that the factors of the quality scan are valid.

7.2 Discussion of research questions

In the previous section, the results of the analyses were discussed and the validity of the quality scan examined. Here, the answers of the research questions will be provided. These research questions were described in chapter three and were divided into four groups. The results will now be presented per group.

7.2.1 Generalisation

The results of the three research questions of this group were for a large part already discussed in chapter three. However, in this section, a short description of the answers will be provided. The first research question concerned the usability of the quality scan in the Dutch Primary Education. It was already argued that the education processes can be divided into several aspects, since this was also done by the Dutch Education Authority. Moreover, the empirical study proved that it was possible for the teachers to answer the different statements. Therefore, it can be concluded that the quality scan can be used in the Netherlands, at least in the district of Zuidoost Brabant.

Second, it was questioned how many adaptations should be necessary before the questionnaire of Marzano (2003) was applicable in the Netherlands. As aforementioned, 44 percent of the statements

in the current quality scan had no direct relation with the statements of Marzano's questionnaire. There are several reasons for the adopted changes in the quality scan. Nine percent of the changes were made, because statements were not relevant for the Primary Education. These were included in Marzano's questionnaire, since his questionnaire was also meant for the Secondary Education. Moreover, it would have been better for Marzano's questionnaire if he made this distinction too, because now both type of education faced several irrelevant statements. The differences between the United States and Dutch education produced 20 percent of the changes. For example, the differentiation of education for students is an important topic in the Netherlands. This reason did not indicate whether the quality of Marzano's questionnaire was good or bad, since his goal was not to develop an international questionnaire. A different reason for the implementation of some statements was that they increased the specification of a factor (i.e. 13 percent). Another 18 percent of the changes were caused by the double meaning of the statements. Several statements of Marzano's questionnaire had a double meaning that could result in bias of the data. Therefore, these statements were split up or deleted. The largest part of the changes was made due to the inclusion of the factors leadership and (teacher) motivation (i.e. 31 percent). Both factors were not only important in the Netherlands, but also in the United States. Thus, the exclusion of these factors in Marzano's questionnaire can be seen as a drawback, since it indicates the incompleteness of the questionnaire. The last 9 percent of the changes were a result of the deletion of four statements, because they were poor stated without a clear content. Therefore, I conclude that Marzano's book provided clear evidence of the importance of the presented factors, but the conversion to statements was not adequately executed.

The third research question concerned how leadership and teacher motivation should be included in the quality scan. As described before, additional literature was consulted. For the category *leadership* some statements of the Multifactor Leadership Questionnaire of Bass and Avolio (2004) were included. These statements had proved to be effective measurements of leadership in a school environment in the Netherlands (Geijsel et al., 1999). The statements of *teacher motivation* were based on several statements of the Job Diagnostic Survey (Hackman & Oldham, 1980) and on statements of a Master Thesis conducted earlier at the Eindhoven University of Technology (Steehouwer, 2007).

7.2.2 Properties of the quality scan

Research question four, the first question of this group, focused on the reliability of the categories and factors of the quality scan. The reliability of the quality scan was measured with the Cronbach's Alpha. For the categories used in the quality scan, only the teacher part of the category *collegiality and professionalism* had a value below 0,60, indicating an inadequate reliability. However, it should be noted that the Cronbach's Alpha is generally higher if there are more items included. This could have affected some categories, since these consisted of a large number of items (Hair et al., 2006). Moreover, not each category formed exactly one factor. Therefore, the reliability of the factors was also measured. The factors *student care system* and *personal development* had a value of 0,59 and 0,51 respectively, which are below the critical value of 0,60. Nevertheless, as described in the previous section these will be used, since the reliability might increase if a larger sample is available.

The fifth research question is about the structure (correlations) among the categories. These correlations were measured for the factors instead of the categories, since these presented a more realistic structure of the quality scan. It followed that there were only positive and in general very high correlations between the different factors. This characteristic causes that an improvement plan

will not affect just one factor, but that it probably has an effect on several factors. A benefit is that an improvement will only affect the factors positively, since all correlations were positive.

7.2.3 Relation with other instruments

The sixth research question is concerned with the EFQM-model. It examined whether the factors can be ascribed to the criteria of the EFQM-model. In chapter six, this classification is provided. The classification followed after consulting the descriptions of the different criteria in literature (Wongrassamee, 2003; Nederlandse Kwaliteit, 1994) and a meeting with an advisor of OCGH Advies. Most of the factors were ascribed to the *processes* criterion as expected, since the quality scan measured the education processes of a school. In addition, some were sorted into the *leadership*, *people management*, *policy and strategy*, *people satisfaction* or *customer satisfaction* criterion. Moreover, 84 percent of the factors were classified under the *enablers* criteria. Hence, the quality scan focuses on the organizational aspects and not on the results of the organizational aspects. This is a desirable outcome, because the goal was to develop and design a quality scan that includes educational related factors influencing the performance of students. Thus, the quality scan should determine the educational aspects at a school. The student performance is the results of these organizational aspects.

Research question seven focused on whether the factors could be related to indicators of the Dutch Education Authority. First, it was attempted to relate the statements of the quality scan with the indicators, but there was too much room for different interpretations. Therefore, this did not result in clear relations between the statements and indicators. Next, an attempt was made to calculate correlations between factors of the quality scan and factors of the Dutch Education Authority. However, there was a lack of variability in the factors of the Dutch Education Authority, so this was no option either. The only conclusion made was that both the quality scan and the Dutch Education Authority revealed the homogeneity of the different schools.

7.2.4 Prediction of student performance

The eighth research question asked what the relations are between the factors and quantitative indicators and measurements of educational effectiveness of a school. For these quantitative measurements, only measures of student performance generalized to the school level were used, since other types of measures were not available (for all schools). In general, the factors had positive relations with student performance. The explanations for negative relations between a few factors and the performance of students were presented in the previous section. Therefore, it was concluded that the factors generally had a positive influence on the performance measures.

Finally, the ninth research question focused on what percentage of the variance was caused by the factors at the school, teacher, and student level respectively, since an indication was provided by Marzano (2003). Unfortunately, this study has not provided an answer to this research question. Due to the small sample size, it was not possible to conduct a regression analysis that should result in the regression variate of each dependent variable (i.e. factor). It can only be argued that 58 percent of the significant positive correlations had to do with factors at the teacher level and 42 percent at the school level. This indicated that there are more relations between the factors at the teacher level and student performance than at the school level. This is in line with the results of Marzano, where it was indicated that 13,3 percent and 6,7 percent of the student performance was predicted by the teacher and school level factors respectively. Thus, the number of relations assumed that the teacher level

factors are stronger related to student performance, but the weight of these relations cannot be established in this study.

7.3 Research limitations

This study like each study had some research limitations. These limitations followed from the choices that were made during the study for data collection, the measurement scales, sampling design, type of analysis, etcetera. The most important research limitations will be described in this section.

First, the sample size was a source of some limitations. Despite the help of several advisors of OCGH Advies, the mailing of a letter to 139 schools to ask whether they were interested to participate, and a call from me to all as reminder, just nine schools participated. The combination of 137 respondents and 88 statements made it impossible to conduct a factor analysis (both exploratory and confirmatory) that included all statements at once. Therefore, I chose to divide the statements in six groups. A result is that not all combinations of statements have been analyzed. Nevertheless, it can be assumed that the resulted factors are adequate, since the statements that were most related were included in the same group. The second problem occurring due to the small sample size was the limited options for the analyses with the performance measures. For these analyses, only a sample size of nine could be used, since all performance measures were at the school level. As a consequence, only correlations with non-parametric statistics could be calculated (Cooper & Schindler, 2003). It was not possible to indicate which factor caused the largest variance to the dependent variable, because of a Multiple Regression Analysis with nineteen factors needs at least 100 observations (Hair et al., 2006).

A second limitation is caused by the time dimension of the study. This was a cross-sectional study; the questionnaire data was gathered at one time and thus represents information of one point in time. Besides, the data of the performance measures were related to a period before the quality scan was completed by the school team. Since the data for the performance measures were not created during the study, but were already available at the school, it was not possible to influence the timing of data collecting. Nevertheless, the most recent available data of the performance measures were used, but no conclusions could be made about the predictability of the quality scan related to the performance measures. In the ideal situation, a longitudinal study should have been conducted. The data of the quality scan and the performance measures should be collected at time one. At time two, the data of the performance measures should be collected again. With this new information, it would be possible to indicate the effect of the quality scan at time one with the performance measures at time two that are corrected for the performance values at time one (Cooper & Schindler, 2003).

Third, the homogeneity of the nine schools was a limitation. Since the quality scan was conducted at homogeneous schools, this has a disadvantage on the generalisation of the conclusions. The schools in this study were not representative for all schools in the Netherlands, but only represented a subpopulation of the schools. These are schools that mainly have students with Dutch parents. However, most of the schools in the district of OCGH Advies have this type of students. Therefore, the study provided useful information for OCGH Advies.

Finally, the type of dependent variables was a limitation for this study. The dependent variables were all types of average student performance at the school level. Therefore, no conclusions could be made concerning the affects at the teacher- or class level. If this would have been the case, the differences between teachers of one school could have been analyzed. Besides, there are more relevant dependent variables for a school than the student performance. The parent satisfaction and employee satisfaction

are examples of relevant dependent variables that have not been used, since there was no available information concerning these variables. Both types are also included in the EFQM model. Nevertheless, the factors of the quality scan have been tested with the most important dependent variable.

7.4 Future research directions

In the previous section, the research limitations of this study have been described. These research limitations can lead to some future research directions. Some of the research directions are scientifically focused, while other research directions can be seen as recommendations for OCGH Advies.

A first follow up study should be carried out with more participating schools. The focus of this study should again be about the structure of the statements and the relations with performance measures. With a larger and a more heterogeneous sample at the school level, an aggregation analysis should be carried out using the interclass correlation coefficient. In addition, a factor analysis at the school level should be conducted, since a factor at one level is not automatically a factor at the higher level (Ostroff & Rothausen, 1997). These analyses should study whether the statements and factors of this study, that were invalid with respect to a criterion, such as the Cronbach's Alpha or the factor loading, could have provided valid results when a larger sample was used. In addition, more performance measures should be used to indicate the relevance of the quality scan at more issues. Examples of performance measures are student satisfaction, parent satisfaction, and employee satisfaction. An additional characteristic of the future study can be to split up the sample and use two different types of statements wording. One half will complete the quality scan with the direct consensus model for the statements at the teacher level, and the other half will have statements with the referent-shift consensus model. With this characteristic, it can be specified whether the statements wording influences the within-group- and between-group variability, and therefore the opportunity to aggregate the factors to the school level.

Another direction is to zoom in at some factors of the quality scan. For example, if the goal is to improve the instructional strategies of a teacher, the statements of this factor should be made more specific. This should result in the development of a project for the teacher so he or she can focus on his or her specific needs. The current quality scan provides the first indications, but it can be extended in more detail.

Third, the predictive validity should be analyzed with a longitudinal study. The design of the study has been described in the previous section. This research design can provide casual relations between the factors and performance measures. However, there are some disadvantages of longitudinal studies. In general, a longitudinal study is quite costly to conduct (Cooper & Schindler, 2003). Subsequently, for this specific study, performance measures should be developed that will be used at all participating schools. The nationwide system of tests is not used at all schools and even when a school uses the system, not all tests are completed. Therefore, it is necessary for OCGH Advies to make clear arrangements with the schools, about which tests should be completed when and how. Nevertheless, the study can provide relevant information and it can increase the commitment of a school to OCGH Advies.

Finally, a recommendation for OCGH Advies can be to include a benchmark in the feedback report. With nine schools, the current benchmark has not yet a value, since it is still very sensitive to the

inclusion or exclusion of one or more schools. However, with more schools (about 25), the benchmark will be a consistent value. Nevertheless, providing a benchmark also gives rise to risks. A school that scores above the benchmark on some factors may conclude that they are performing well (enough). Therefore, they will not try to improve the aspects related to these factors, while these aspects could provide important improvements. Thus, a benchmark has both advantages and disadvantages. As aforementioned, I recommend not to include a benchmark in the feedback report. However, the benchmark should be calculated, so the advisors have some additional information available to put the results of the statements and factors into perspective. This information can be used when the report is being discussed with a school.

7.5 Conclusion

In this chapter, the results have been discussed, the research limitations provided, and possible future research directions were indicated. In this section, some concluding comments will be made concerning this study.

First, it can be concluded that an adequate literature study was very helpful as starting point for my Master Thesis project. During this literature study, insights in the relevant aspects of the Primary Education were presented. I found out that Marzano's questionnaire (2003) did not provide all relevant statements. Nevertheless, during only a literature study, it is not possible to gather all necessary information. Therefore, the direct information, gathered by interviews with advisors of OCGH Advies, was also valuable. As a result, the first part of the project objective, the design of a quality scan, was conducted.

The second part of the project objective, the validation of the quality scan, has been completed with an empirical study. The analyses of this study offered evidence to conclude the presence of different types of validity. The interval validity of the quality scan was supported by presenting the presence of face, convergent, nomological, and concurrent validity. The external validity was limited to a subpopulation of the primary schools of the Netherlands, due to the homogeneous sample. However, most schools in the region of OCGH Advies belong to this subpopulation.

Concluding, both parts of the project objective have been accomplished. In addition, the quality scan has been adjusted based on the results of the validation. The statements that statistically did not belong to a factor were included in the quality scan based on their content. Therefore, the quality scan was scientifically tested and finally the practical relevance was encountered. Thus, a final conclusion is that the quality scan can be valuable for OCGH Advies.

8 References

Avolio, B.J. (1999). *Full leadership development: building the vital forces in organizations.* London; Saga.

Babad, E., Avni-Babad, D. and Rosenthal, R. (2003). Teachers' brief nonverbal behaviors in defined instructional situations can predict students' evaluations. *Journal of Educational Psychology*, Vol. 95, No. 3, pp. 553-562.

Balbinotti, M.A.A., Benetti, C. and Terra, P.R.S. (2006). Translation and validation of the Graham-Harvey survey for the Brazilian context. *International Journal of Managerial Finance*, Vol. 3, No. 1, pp. 26-48.

Bass, B.M. and Avolio, B.J. (2004). *Multifactor Leadership Questionnaire: Third edition manual and sampler set.* Mind Garden, Inc.

Biemiller, A. and Boote, C. (2006). An effective method for building meaning vocabulary in primary grades. *Journal of Educational Psychology*, Vol. 98, No. 1, pp. 44-62.

Bliese, P.D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K.J. Klein & S.W.J. Kozlowski (Eds.), *multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349-381). San Francisco: Jossey-Bass.

Butler, R. (2007). Teachers' achievement goal orientations and associations with teachers' help seeking: examination of a novel approach to teacher motivation. *Journal of Educational Psychology*, Vol. 99, No. 2, pp. 241-252.

Butterfield, K.D., Trevino, L.K. and Ball, G.A. (1996). Punishment from the manager's perspective: a grounded investigation and inductive model. *Academy of Management Journal*, Vol. 39, No. 6, pp. 1479-1512.

Chan, D. (1998). Functional relations among constructs in the same content domain at the different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, Vol. 83, No. 2, pp. 234-246.

Cooper, D.R. and Schindler P.S. (2003). *Business research methods (eighth edition).* New York, VA: McGraw-Hill.

DiPerna, J.C., Lei, P.W. and Reid, E.E. (2007). Kindergarten predictors of mathematical growth in the primary grades: an investigation using the early childhood longitudinal study-kindergarten cohort. *Journal of Educational Psychology*, Vol. 99, No. 2, pp. 369-379.

Ehrlich, C. (2006). The EFQM-model and work motivation. *Total Quality Management,* Vol. 17, No. 2, pp. 131-140.

Englund, M.M., Luckner, A.E., Whaley, G.J.L. and Egeland, B. (2004). Children's achievement in early elementary school: longitudinal effects of parental involvement, expectations, and quality of assistance. *Journal of Educational Psychology*, Vol. 96, No. 4, pp. 723-730.

Epstein, J.L. (2005). Links in a professional development chain: preservice and inservice education for effective programs of school, family, and community partnerships. *The New Educator*, 1:2, pp. 125-141.

Fagan, W.T. (1998). Cross-generational perspectives on educational issues: curriculum implications. *Curriculum Enquiry*, Vol. 28, No. 1, pp. 9-25.

Gamero, N., González-Romá, V. and Peiró J.M. (2008). The influence of intra-team conflict on work teams' affective climate: A longitudinal study. *Journal of Occupational and Organizational Psychology*, Vol. 81, pp. 47-69.

Geijsel, F., Sleegers, P. and Van Den Berg, R. (1999). Transformational leadership and the implementation of large-scale innovation programs. *Journal of Educational Administration,* Vol. 37, No. 4, pp. 309-328.

Griffith, S.A. (2008). A proposed model for assessing quality of education. *International Review of Education*, Vol. 54, pp. 99-112.

Hackman, J.R. and Oldham, G.R. (1980). *Work redesign*. Massachusetts, VA: Addison-Wesley Publishing Company.

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E. and Tatham, R.L. (2006). *Multivariate data analysis (sixth edition)*. New Jersey, VA: Pearson Prentice Hall.

Hughes, J.N., Luo, W., Kwok, O.M. and Loyd, L.K. (2008). Teacher-student support, effortful engagement, and achievement: a 3-year longitudinal study. *Journal of Educational Psychology*, Vol. 100, No. 1, pp. 1-14.

Inspectie van het Onderwijs (2005). *Toezichtkader primair onderwijs 2005*, Utrecht.

Klein, K.J., Bliese, P.D., Kozlowski, S.W.J., Dansereau, F., Gavin, M.B., Griffin, M.A., et al. (2000). Multilevel analytical techniques: Commonalities, differences and continuing questions. In K.J. Klein & S.W.J. Kozlowski (Eds.), *multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349-381). San Francisco: Jossey-Bass.

Klein, K.J., Conn, A.B., Smith, D.B. and Sorra, J.S. (2001). Is everyone in Agreement? An exploration of within-group agreement in employee perceptions of the work environment. *Journal of Applied Psychology*, Vol. 86, No. 1, pp. 3-16.

Koth, C.W., Bradshaw, C.P. and Leaf P.J. (2008). A multilevel study of predictors of student perceptions of school climate: the effect of classroom-level factors. *Journal of Educational Psychology*, Vol. 100, No. 1, pp. 96-104.

Lapadat, J.C. (2002). Relationships between instructional language and primary students' learning. *Journal of Educational Psychology*, Vol. 94, No. 2, pp. 278-290.

Lau, S. and Nie, Y. (2008). Interplay between personal goals and classroom goal structures in predicting student outcomes: a multilevel analysis of person-context interactions. *Journal of Educational Psychology*, Vol. 100, No. 1, pp. 15-29.

Leenders, Y., Meyer, P., Sanders, M. and Veenman, S (1993). *Effectieve instructie: leren onderwijzen met behulp van het directe-instructiemodel.* CPS, Hoevelaken.

Lemos, M.S. (1999) Students' goals and self-regulation in the classroom. *International Journal of Educational Research*, Vol. 31, No. 6, pp 471-485.

Li, Y. (2002). Knowing, understand and exploring the content and formation of curriculum materials: a Chinese approach to empower prospective elementary school teachers pedagogically. *International Journal of Educational Research*, Vol. 37, pp. 179-193.

Lovelace, K., Shapiro, D.L. and Weingart, L.R. (2001). Maximizing cross-functional new product teams' innovativeness and constraint adherence: a conflict communications perspective. *Academy of Management Journal,* Vol. 44, No. 4, pp. 779-793.

Marzano R.J. (2003). *What works at schools: translating research into action*. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).

Nederlandse Kwaliteit (1994). *Handleiding positiebepaling & verbetering ondernemingen: op weg naar de Nederlandse kwaliteitsprijs.* 's-Hertogenbosch, instituut Nederlandse Kwaliteit.

Ostroff, C. and Rothausen, T.J. (1997). The moderating effect of tenure in person-environment fit: A field study in educational organizations. *Journal of Occupational and Organizational Psychology,* Vol. 70, pp. 173-188.

Pintrich, P.R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*, Vol. 95, No. 4, pp. 667-686.

Porter, L.W., Bigley, G.A. and Steers, R.M. (2003) *Motivation and work behavior.* New York, VA: McGraw-Hill.

Pressley, M., Mohan, L., Raphael, L.M. and Fingeret, L. (2007). How does bennett woods elementary school produce such high reading and writing achievement? *Journal of Educational Psychology*, Vol. 99, No. 2, pp. 221-240.

Reeve, J. and Jang, H. (2006). What teachers say and do to support students' autonomy during a learning activity. *Journal of Educational Psychology*, Vol. 98, No. 1, pp. 209-218.

Reijn, G. (2007). Basisonderwijs presteert ondermaats. *Volkskrant,* 23 april 2007.

Roth, G., Assor, A., Kanat-Maymon, Y. and Kaplan, H. (2007). Autonomous motivation for teaching: how self-determined teaching may lead to self-determined learning. *Journal of Educational Psychology*, Vol. 99, No. 4, pp. 761-774.

Ryan, R.M. and Deci, E.L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist,* Vol. 55, No. 1, pp. 68-78.

Scandura, T.A. and Williams, E.A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management Journal*, Vol. 43, No. 6, pp. 1248-1264.

Schmitt, N., Sacco, J.M., Ramey, S., Ramey, C. and Chan, D. (1999). Parental employment, school climate, and children's academic and social development. *Journal of Educational Psychology*, Vol. 84, No. 5, pp. 737-753.

Sideridis, G.D. (2005). Classroom goal structures and hopelessness as predictor of day-to-day experience at school: differences between students with and without learning disabilities. *International Journal of Educational Research*, Vol. 43, No. 4-5, pp. 308-328.

Snijders T.A.B. and Bosker R.J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, VA: Sage.

Steehouwer, H.G. (2007). The translation and validation of the Dutch version of the motivation assessment questionnaire; final report. *Afstudeerverslag Technische Universiteit Eindhoven*, ARW 2007 BDK 4679.

Terugblik en resultaten, 2008. Eindtoets basisonderwijs groep 8. *Versie 29 februari 2008.*

Trautwein, U. and Lüdtke, O. (2007). Students' self-reported effort and time on homework in six school subjects: between-students differences and within-student variation. *Journal of Educational Psychology*, Vol. 99, No. 2, pp. 432-444.

Werkgroep Opbrengsten PO (2008). Analyse en waarderingen van opbrengsten: primair onderwijs. *Utrecht, september 2008.*

Wolters, C.A. and Daugherty, S.G. (2007). Goal structures and teachers' sense of efficacy: their relation and association to teaching experience and academic level. *Journal of Educational Psychology*, Vol. 99, No. 1, pp. 181-193.

Wongrassamee, S., Gardiner, P.D. and Simmons, J.E.L. (2003). Performance measurement tools: the balanced scorecard and the EFQM excellence model. *Measuring Business Excellence,* Vol. 7, No. 1, pp. 14-29.

Appendix A: Secondary Education

The types of Secondary Education for each group are presented in Table B1. This division was chosen after consulting an advisor of OCGH Advies. The distinction between the different VMBO levels is not done the same for each Secondary Education school. Therefore, these two types of the VMBO were taken together. In addition, not every Primary Education school made a distinction between the VWO and Gymnasium, which resulted in the combination of these two levels.

Group	**Level of Secondary Education**
Group 1	Praktijk Onderwijs
Group 2	VMBO Kader (met en zonder LWOO) VMBO Basisberoepsgericht (met en zonder LWOO)
Group 3	VMBO gemengd (met en zonder LWOO) VMBO theoretisch (met en zonder LWOO)
Group 4	VMBO / Havo
Group 5	Havo
Group 6	Havo / VWO
Group 7	VWO Gymnasium

Table B1: The level of Secondary Education for each group.

Appendix B: Calculation LVS variables

The class results of a LVS is described in the percentage of students having an A,B,C,D and E value respectively. Here, an A value is the highest and an E value the lowest. The test scores were collected for the last two years. For each class an average score of a test could be calculated per year. This was done with the equation:

$$Value_{ij} = 5 * percent\ A + 4 * percent\ B + 3 * percent\ C + 2 * percent\ D + 1 * percent\ E$$

With i = current class number {3,4,5,6,7,8}
j = school year {2006 / 2007 ; 2007 / 2008}

For the variables *LVS average*, the next equation was:

$$LVS\ average = \frac{1}{total\ number\ of\ values} \sum \frac{Value_{ij}}{100}$$

This equation resulted in a value between one and five. For the *LVS delta* the difference between the class average of school year 2007 / 2008 and the class average of school year 2006 / 2007 for each class was taken. The school value of this variable was the average of all class variables.

Appendix C: EFA of group 2

In section 5.3.2, the exploratory factor analysis process of group two is described. There were two exploratory factor loading matrixes not presented. These are given in this Appendix. In Table D1, the EFA without item SM30 is provided, and in Table D2, the EFA without items SM30 and PI22 is presented.

Item	**Component**		
	1	2	3
SM27 Feedback on knowledge increase	0,84		
SM28 Meaningful learning activities	0,77		
SM29 Self invented projects	0,49		
SM31 School wide education	0,48		
PI22 Helping parents by raising their child			
PI20 Low barrier for parents to contact school		0,80	
PI21 Parental involvement in daily activities		0,78	
PI19 Contact teachers with parents		0,68	
SE25 Consequence violations of school rules			-0,83
SE24 Clear school rules formulated			-0,79
SE23 Promotion of good behaviour			-0,56
SE26 Improvement of self-discipline			-0,49
Eigenvalue	2,74	2,46	2,70

Notes: SM = student motivation, PI = parent and community involvement, SE= safe and orderly environment. Extraction = principal component analysis; rotation = Direct Oblimin.

Table D1: Exploratory factor analysis of group 2 without item SM30.

Item	**Component**		
	1	2	3
SM27 Feedback on knowledge increase	0,84		
SM28 Meaningful learning activities	0,78		
SM29 Self invented projects	0,48		
SM31 School wide education	0,48	0,40	
PI20 Low barrier for parents to contact school		0,79	
PI21 Parental involvement in daily activities		0,78	
PI19 Contact teachers with parents		0,68	
SE25 Consequence violations of school rules			-0,83
SE24 Clear school rules formulated			-0,80
SE23 Promotion of good behaviour			-0,56
SE26 Improvement of self-discipline			-0,49
Eigenvalue	2,50	2,30	2,46

Notes: SM = student motivation, PI = parent and community involvement, SE= safe and orderly environment. Extraction = principal component analysis; rotation = Direct Oblimin.

Table D2: Exploratory factor analysis of group 2 without items SM30 and PI22.

Appendix D: CFA of group 1

In section 5.3.3, the process of the confirmatory factor analysis (CFA) of group one is formulated. This Appendix provides background information about this process. In Table E1, the Lisrel syntax, which established the model of Table 5.9, is presented.

```
da no=137 ni=17
ra fi='D:\My Documents\Afstuderen\Analyse\Group 1\FACTORS GROUP 1.psf'
la
GC7 GC8 GC9 GC10 GF11 GF12 GF13 GF14 GF15 GF16 GF17
GF18 CP_G42 CP_G43 CP_G44 CP_G45 CP_G46 /
se
GC7 GC8 GC10 GF11 GF12 GF13 GF14 GF15 GF16 GF17
GF18 CP_G42 CP_G44 CP_G45 CP_G46 /
mo nx=15 nk=5 lx=fu,fr ph=st,sy td=di,fr
pa lx
3(1 0 0 0 0)
2(0 1 0 0 0)
4(0 0 1 0 0)
2(0 0 0 1 0)
4(0 0 0 0 1)
lk
Cur SCS L_goals D_goals G_prof
pd
ou nd=4 mi sc
```

Table E1: Lisrel Syntax for CFA group 1.

The results of the model without the factor *student care system* are provided in Table E2.

Factor	**Item**		**Loadings**	**t-value**
Fit Indices	Chi-Square = 96,90 (d.f. = 59, p-value = 0,0014); RMSEA = 0,057; CFI = 0,959; GFI = 0,912; AGFI = 0,865			
1 *Education curriculum*				
	GC7	Knowledge of essential content	0,55	6,08
	GC8	Organisation of content	0,80	8,99
	GC10	Realisation effective learning time	0,65	7,29
2 *Class learning goals*				
	GF13	Individual schedules with goals	0,72	8,86
	GF14	Individual schedules used for actions	0,80	10,22
	GF15	Group schedules with goals	0,66	8,04
	GF16	Group schedules used for actions	0,67	8,12
3 *School development goals*				
	GF17	Formulation of school development goals	0,84	9,94
	GF18	Development goals used for policy plans	0,87	10,27
4 *Team professionalism*				
	CP_G42	Norms for professionalism	0,75	8,19
	CP_G44	Providing constructive feedback	0,61	6,64
	CP_G45	Be open to receive feedback	0,52	5,53
	CP_G46	Stand up for myself	0,29	2,93

Note: t-value > 1,96 indicates that the factor loading is significant at p<0,05.

Table E2: Confirmatory factor analysis group 1 without factor "student care system".

Finally, the different model fits for each group are provided in Table E3.

Model	**Chi-Square (d.f. ; p-value)**	**RMSEA**	**CFI**	**GFI**	**AGFI**
Group 1	136,53 (80 ; 0,0001)	0,063	0,946	0,890	0,840
Group 2 (with SM29)	86,57 (41 ; 0,0000)	0,090	0,907	0,897	0,834
Group 3	182,41 (76 ; 0,000)	0,096	0,962	0,847	0,789
Group 4	157,43 (94 ; 0,000)	0,067	0,940	0,871	0,813
Group 5	145,89 (87 ; 0,0001)	0,074	0,939	0,850	0,793
Group 6	-	-	-	-	-

Note: d.f. = degrees of freedom

Table E3: The different fit indices for the model of each group.