Value added in educational accountability

Timmermans, A.C.

Value added in educational accountability: Possible, fair and useful?

Anneke Timmermans

university of groningen / faculty of behavioural and social sciences

ico

ISBN printed version: 978-90-367-5850-5
ISBN electronic version: 978-90-367-5852-9
Cover illustration: Ubel Smid
Printing: GVO drukkers & vormgevers B.V. | Ponsen & Looijen

RIJKSUNIVERSITEIT GRONINGEN

Value added in educational accountability: Possible, fair and useful?

Proefschrift

ter verkrijging van het doctoraat in de
Gedrags- en Maatschappijwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
dinsdag 27 november 2012
om 11.00 uur

door

Antje Cobie Timmermans
geboren op 2 augustus 1983
te Leeuwarden

Promotor:	Prof. dr. R.J. Bosker
Co-promotoren:	Dr. S. Doolaard Dr. I.F. de Wolf
Beoordelingscommissie:	Prof. dr. W.J.C.M. van de Grift Prof. dr. F.J.G. Janssens Prof. dr. S. Thomas

Voorwoord

In de vier jaar dat ik aan dit proefschrift heb mogen werken zijn er een aantal mensen heel belangrijk voor mij geweest en die ik graag wil bedanken. In de eerste plaats gaat mijn dank uit naar mijn promotor Roel Bosker en mijn dagelijks begeleiders Simone Doolaard en Inge de Wolf. Jullie hebben mij alle vrijheid gegeven om mijn eigen richting te geven aan het onderzoek. Roel, bedankt voor je kritische opmerkingen en je vermogen om de resultaten in een grotere context te zetten. Simone, dank je wel voor al je steun, je goede tips, maar vooral je hulp bij het vertalen van de resultaten van de ingewikkelde analyses naar 'gewone mensen taal'. Inge, bedankt voor je enthousiasme, je aanmoedigingen voor het verbeteren van het huidige onderzoek en ontwikkelen van nieuwe ideeën. Ik heb je gemist toen je in het verre Amerika was.

Tevens wil ik alle GION collega's bedanken. Hanke, Anouk en Mechteld, bedankt voor "logisch nadenk momentjes", wanneer ik weer eens resultaten vond die op het eerste gezicht niet heel logisch leken. Truus, bedankt voor de mogelijkheid om mee te werken aan je project op het Noorderpoort college en het meeschrijven aan een artikel los van dit proefschrift. Alma, ik mocht altijd even gezellig komen bijkletsen. Dank je wel. Jolijn en Lyset, we hebben de afgelopen jaren veel samengewerkt bij het geven van onderwijs. Ik heb veel van jullie geleerd.

Alle collega's van de Inspectie van het Onderwijs wil ik graag bedanken, ook al kwam ik vaak maar één dag in de week en sloeg ik ook wel eens een paar weken over. In het bijzonder wil ik Annet en Machteld en als bijzonder prettige kamergenoten. Maar ook Margriet, Maarten en Ineke, bedankt voor alle goede gesprekken over het wel en wee van het promoveren. Geertje en Bruno, voor de bijzondere samenwerking in verschillende werkgroepen over onderwijsopbrengsten en toegevoegde waarde. En alle andere kenniscollega's en inspecteurs die in min of meerdere mate betrokken waren bij toegevoegde waarde.

Leden van de ICO-onderwijscommissie en ICO-Themagroep (Adrie en Sjoerd), bedankt voor de fijne samenwerking en de afleiding die het gaf van het "gewone"

werk. Sally and George, thank you for your enthusiasm, the opportunity to visit you in Bristol and the possibility to present and discuss my research.

Ten slotte wil ik mijn familie en vrienden bedanken. Heit en Mem en Menno, jullie hebben mij altijd door dik en dun gesteund. Loty, Bianca, Jan en Alexandra bedankt voor alle afleiding en gezellige momenten. Danny, ik leerde je kennen in het derde jaar van mijn promotie traject toen je bij ons op het GION kwam werken. Je bent mij heel dierbaar geworden en hebt mij veel geholpen tijdens de makkelijke en minder makkelijke momenten van het promoveren.

Anneke Timmermans

Contents

Chapter 1 **11**
Introduction

1.1 Introduction 12

1.2 Educational accountability 12

1.3 Current performance indicators in Dutch educational accountability 14

1.4 Value added 17

1.5 The translation of the concept of value added into statistical models 18

1.6 Validity and reliability of value added 21

1.7 Methodological challenges in modelling value added 24

1.8 The present dissertation 26

Chapter 2 **29**
Conceptual and Empirical Differences among Various Value Added Models for Accountability

2.1 Introduction 31

2.2 Method 36

2.3 Results of the analysis 42

2.4 Conclusion and discussion 49

Chapter 3 **55**
In Search of Value Added in the Case of Complex School Effects

3.1 Introduction 57

3.2 Method 62

3.3 Results 67

3.4 Conclusion and discussion 71

Chapter 4 **75**
Educational Accountability Based on the Cognitive and Non-cognitive Performance of Students: A Value Added Approach

4.1 Introduction 77

4.2 Method 81

4.3 Results 87

4.4 Conclusion and discussion 94

Chapter 5 **97**
Value Added Based on Educational Careers in Dutch Secondary Education

5.1 Introduction 99

5.2 Method 104

5.3 Results 111

5.4 Conclusion and discussion 121

Chapter 6 **125**
Value Added as an Indicator of Educational Effectiveness in Dutch Senior Secondary Vocational Education

6.1 Introduction 127

6.2 Method 130

6.3 Results 132

6.4 Conclusion and discussion 141

Chapter 7 **145**
Risk-based Educational Accountability in Dutch Primary Education

7.1 Introduction 147

7.2 Method 150

7.3 Results 159

7.4 Conclusion and discussion 173

Chapter 8 **177**
General Discussion

8.1 Main findings 178

8.2 Theoretical and practical implications 182

8.3 Limitations 186

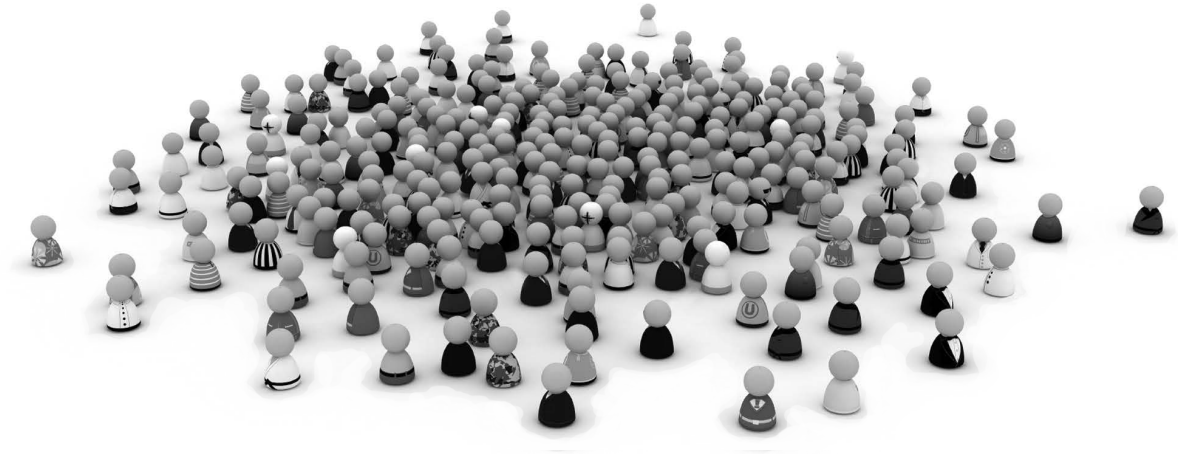
8.4 Future directions 188

Nederlandse Samenvatting 191

References 203

ICO Dissertation Series 219

Chapter 1
Introduction



1.1 Introduction

For an accurate identification of underperforming schools, any governmental body with accountability tasks, such as the Inspectorate of Education, needs reliable and valid indicators of school performance based on achievement of students. Since 1997, explorations have been conducted in the Netherlands with respect to the development of value added, an indicator of the effectiveness of schools, based on student level performance data (Bosker, Lam, Dekkers, & Vierke, 1997; Bosker, Lam, Luyten, Steen, & Vos, 1998; Bosker, Béguin, & Rekers-Mombarg, 2001; Inspectie van het Onderwijs, 2003; Verhelst, Staphorsius, & Kleintjes, 2003; Wijnstra, Ouwens, & Béguin, 2003; Roeleveld, 2003a; Van de Grift, 2009). Some of these explorations show a more specific focus on accountability. Several of these explorations corroborate the importance of value added as a part of a set indicators for the quality of schools (Onderwijsraad, 2003; Inspectie van het Onderwijs, 2003; Verhelst et al., 2003; Wijnstra et al., 2003). In this dissertation, the use of value added for educational accountability is explored in more detail. The main aims of the studies conducted in this dissertation are a) to develop value added models for school within the context of educational accountability, b) to study whether or not using value added in educational accountability would lead to valid comparisons of the performance of schools, and c) whether value added can be used in a risk based educational accountability system to predict future underperformance of schools.

The first part of this chapter starts with a description of the Dutch educational accountability system and the role of the Inspectorate of Education, the context of the studies conducted in this dissertation. The next part of this chapter will provide a description of the current performance indicators in Dutch educational accountability. After that, the concept of value added as indicator of school quality is introduced and a discussion on the validity, reliability and methodological challenges of value added will be presented. In the final paragraph an overview of the dissertation is given in which several aspects of value added are highlighted.

1.2 Educational accountability

In the Netherlands, the Dutch Education Supervision Act (2012) describes the tasks and the formal position of the Inspectorate of Education and prescribes the global

framework for education accountability. The first task of the inspectorate is to judge the quality of education by means of research of compliance with the law on education and several quality aspects as described in the supervision act (Ehren, De Leeuw, & Scheerens, 2005). Educational laws describe several requirements for schools as a prerequisite for receiving governmental funding. Because the legal requirements alone are considered to be insufficient, the Educational Supervision Act further specifies quality standards (that partly elaborate on the legal requirements). In the Dutch Education Supervision Act two major categories of educational quality aspects are described, namely quality as measured by educational outcomes of students and the realization of the educational learning process within schools. In this act, educational outcomes are described as student performance and progress in the development of students. The category educational learning process contains aspects of learning time, pedagogical climate, school climate, quality assurance, care for special needs students, testing and examinations and teacher quality.

The Inspectorate of Education translated the Dutch Education Supervision Act in an, by the minister approved, accountability framework based on findings from educational effectiveness research (Inspectie van het Onderwijs, 2006; Inspectie van het Onderwijs, 2009). Both the working method and operationalization of indicators concerning compliance of the law and quality aspects of education are described in more detail in this framework.

This explicitness of the focus on quality has changed during the more than 200 years existence of the Dutch Inspectorate of Education. Tasks have changed from policy development, to cooperation in the implementation of education, to accountability (Mertens, 2002; Mertens, 2009; Inspectie van het Onderwijs, 2011e). The recent changes in tasks and focus on accountability of the Inspectorate of Education are considered to be a product of changing beliefs on the role of the government since the 1980's towards more autonomy for schools (Onderwijsraad, 1999; Elte & Scholtes, 2002; Mertens, 2009; Inspectie van het Onderwijs, 2011e). Similar developments have taken place at the Inspectorate of Healthcare around the same time (Mertens, 2009).

In several other countries accountability systems have a similar focus on both compliance to regulations and quality aspects of education. An example of a fairly similar inspection framework is provided by Ofsted in England, the Office of Standards in Education (Ofsted, 2010; Ofsted, 2011). Like the Dutch accountability framework, the inspection framework of Ofsted is based on findings from educational effectiveness research in the United Kingdom (Sammons, Hillman, & Mortimore,

1995a). The inspection framework describes both indicators of educational outcomes as well as of the learning process and quality assurance.

A recent development in educational accountability in the Netherlands and England is a risk based strategy to improve the efficiency and effectiveness of educational accountability (Inspectorate of Education, 2009; Ofsted, 2011). A risk based strategy implies that the intensity and/or frequency of school inspections can vary across schools depending upon the results of previous inspections and their subsequent performance. Underperforming schools are inspected more and outstanding schools less frequently. Both inspectorates use a methodology in which annually a risk assessment is conducted based on the past and current performance of schools and multiple other signals, for example signals concerning children's safety within schools. This risk assessment determines which school are "at risk" and should be visited in the upcoming year. The annual risk assessment depends heavily on an adequate estimation of the performance of schools and models for estimating possible risks. Furthermore, in both methodologies the risk assessment is followed up by an in depth investigation by inspectors of schools that show possible risks. This whole risk assessment process leads to tailored inspections for each school.

1.3 Current performance indicators in Dutch educational accountability

1.3.1 Primary education

The accountability framework for Dutch primary education contains 5 performance indicators in total (Inspectie van het Onderwijs, 2011b). These performance indicators refer to cognitive achievement at the end of primary education, cognitive achievement during primary education, the amount of grade retention, the performance of students with special educational needs and social competences. Next to the performance indicators, school processes, policy and social outcomes are assessed during school inspections. The performance indicators are based on the results of students on tests at the end or primary education, tests in monitoring systems during primary education and the amount of grade retention. To pursue fair comparisons of the performance of schools at the end of primary education, a comparison is made between a schools' scores on a test and the results of schools with a similar student population. This latter is based on the percentage of students within schools with lowly educated parents, in the Netherlands also known as "gewichtenregeling". For the stability of the indicators, the final judgment on the performance of schools is based on the results of the last three years. This current methodology for estimating the performance of schools is

based on school level data. Separate performance indicators are formulated for students with special educational needs and the performance of students in the social domains. For an extensive overview of the indicators and norms see "Analyse en waarderingen van opbrengsten primair onderwijs" (Inspectie van het Onderwijs, 2011b).

1.3.2 Secondary education

Four performance indicators are defined within the accountability framework for Dutch secondary education (Inspectie van het Onderwijs, 2009; Inspectie van het Onderwijs, 2011c). Similar to primary education, the process and policy within schools is assessed during school inspections. The indicator "efficiency during the first two years of secondary education" is based on 1) a comparison between the primary school advice and the position of students in secondary education at the start of the third year, and 2) the amount of grade retention in the first two years of secondary education (Inspectie van het Onderwijs, 2011d). This indicator is estimated for a complete school and can contain information on multiple school tracks and multiple school locations. The final judgment on the performance of schools is based on the moving averages over the results of the last three years.

The second performance indicator in the educational accountability framework in Dutch secondary education is the efficiency during the final years of secondary education. This indicator is estimated separately for school tracks within schools. This indicator is not based on cohort data, but cross-sectional data on one school year is used to estimate the probability of graduating without grade retention (Inspectie van het Onderwijs, 2011d). For example, in one school year 95% of the students in the theoretical track of pre-vocational education are promoted from third to the fourth grade and 90% of the students in fourth grade graduate. First, the average probability for promotion in third and fourth grade is calculated through (0.95*0.90)/2=0.925. Thereafter the probability is calculated for graduating without grade retention, based on the average probability of promotion, through 0.925*0.925=0.856. This implies that the calculated probability of graduation without grade retention for students in theoretical track of pre-vocational education in this school is 86%. The efficiency of schools is considered insufficient for those schools with the lowest 25% scores on this indicator.

The third indicator for secondary education is a comparison of the examination grades of students between schools. Similar to the previous indicator, the examination

grade indicator is estimated separately for school tracks within schools. A school level regression model is used to control for differences in student populations between schools in order to pursue fair comparisons of the performance of secondary schools (Inspectie van het Onderwijs, 2011d). Based on the percentage of students living in problematic neighbourhoods, students with special educational needs and the percentage of students entering the school in the third grade, a prediction is made for the examination grades. Thereafter, the difference between the predicted grades and the actual grades is calculated. Average schools with higher actual than predicted scores on the examination realize higher grades than might be expected given their student population.

The final indicator in the accountability framework of secondary education is the difference in grades between the central examination and the school examination. School examinations are tests developed by schools that are administered during the final years of secondary education. The final grades of students are based on the grades on the central and school examinations. To prevent diploma inflation, the scores on the school examination should be fairly similar to the scores on the central examinations. School examination scores over a half grade higher than the central examination scores are considered a great difference.

1.3.3 Vocational education

The educational accountability framework for Dutch senior secondary vocational education contains only two efficiency indicators for the performance of institutions (Inspectie van het Onderwijs, 2011f). The two indicators are estimated for the complete educational institutions in vocational education and for clusters of training programmes within institutions, based on the so-called "qualification files" (Inspectie van het Onderwijs, 2011a). These indicators are developed in cooperation with a number of stakeholders in Dutch senior secondary vocational education. Both indicators are based on the number of graduated students and the number of school leavers, however they differ in operationalization. The indicator "jaarresultaat" is based on the number of graduated students in a given year and the number of school leavers without a diploma in a given year. This indicator includes only those students that graduate in the current year. The graduated students in a given year are included whether they left the educational institution or not. The indicator "diplomaresultaat" is the ratio of the number of graduated school leavers in a given year over all school leavers in a given year. This indicator includes all graduated students whether or not they graduated in the current year or before. Both these indicators do not account for

differences in student populations between institutions or clusters of training programmes.

It is questionable whether the indicators, as described above, make a fair comparison between educational institutions possible. Almost all of the performance indicators currently used in Dutch educational accountability are estimated based on school level data. As we will see in the following paragraphs, it is necessary for an accurate estimation of the school effects to take the hierarchical structure of the data into account. Furthermore, in the current performance indicators different methods are used to make the indictors comparable, for example by comparing with similar groups of schools or school level regression analysis. Prior achievement of students is usually ignored in these indicators in making them comparable over schools. In this dissertation, the importance of including prior achievement of student in the estimation of the performance of schools is shown.

1.4 Value added

Value added is originally an economic concept based on the input, energy and output of organizations or companies and has subsequently been introduced in education as a measure of school quality (Saunders, 1999). Over time, several definitions have been given to the concept of value added. For example, "the contribution of a school to student's progress towards stated or prescribed education objectives (e.g. cognitive achievement). The contribution is net of other factors that contribute to students' educational progress" (OECD, 2008, p. 17). Or value added can be defined as "an indication of the extent to which any given school has fostered the progress of all students in a range of subjects during a particular time period in comparison to the effects of other schools in the same sample" (Sammons, Thomas, & Mortimore, 1997). For this dissertation a slightly different definition of value added is used, which is: "Value added is a measure of relative achievement or progress of students in one school compared to students in other schools in the same sample after controlling for differences between students outside the control of the school that influence student achievement."

Literature has shown that there are several other factors than "the school" that can influence the scholastic development of children (Bosker et al., 1998; Teddlie & Reynolds, 2000a; Ten Dam & Vermunt, 2003; Gutman, Sameroff, & Cole, 2003; Doolaard & Leseman, 2008), and these factors therefore influence the difference

between prior academic achievement and final academic achievement and therefore the estimation of value added. Socio-economic status (Willms, 1986; Duncan & Brooks-Gunn, 2000; Bradley & Corwyn, 2002; Ackerman, Brown, & Izard, 2004; Peetsma, Van der Veen, Koopman, & Van Schooten, 2006), ethnicity, gender (Dekkers, Bosker, & Driessen, 2000), level of education of the parents (De Fraine, Van Damme, Van Landeghem, Opdenakker, & Onghena, 2003), general context characteristics of the school (Willms, 1986; Teddlie et al., 2000a; Opdenakker & Van Damme, 2001; De Fraine et al., 2003) and the general context characteristics of the neighbourhood (Leventhal & Brooks-Gunn, 2004) can be seen as important factors that are related to student achievement and are beyond the control of the school. The influence of these factors is unevenly distributed between the schools, as there are large differences in student populations between schools (Hill & Rowe, 1996). Therefore, unadjusted averages of individual performance of students aggregated to the school level are considered insufficient and unfair as an indicator of school performance (Meyer, 1997; Webster, Mendro, Orsak, & Weerasinghe, 1998). To find the unique value added of a school one should incorporate "all" of the factors beyond the control of the school, that influence the development of learning and cognitive abilities, into the statistical analysis to isolate the contribution of a school.

The development of multilevel statistical models to estimate value added (Raudenbush & Bryk, 1986; Aitkin & Longford, 1986; Willms & Raudenbush, 1989; Hill et al., 1996; Goldstein, 1997) caused a rapid development of research into value added. The multilevel models provide the opportunity to analyse which part of the variance in academic achievement is due to differences between students and which part is associated with the school level (Creemers & Sleegers, 2003). Over time a large amount of literature was built up in which value added was mentioned as a reasonable method for estimating the effects of schools (Sammons, Nuttall, & Cuttance, 1993; Mortimore & Sammons, 1994; Meyer, 1997; Onderwijsraad, 2003; Schagen & Hutchison, 2003; OECD, 2008).

1.5 The translation of the concept of value added into statistical models

Several methods have been developed to calculate estimates of value added. The median method is a methodology for the calculation of value added estimates based on a single input and output measure (Tymms & Dean, 2004; Ray, 2006). This method was developed to be simple and understandable for schools, boards and teachers. Due to its simplicity it is possible for schools to calculate their own value added based on national median lines. Based on national data a median for final achievement can be

found for every possible level of prior achievement. This median of final achievement can be seen as the expected result for students given a particular level of prior achievement. The over-/underachievement scores for each individual child can be derived by substracting the expected final achievement of the achieved score. Averaging all the individual over-/underachievement scores of the students within a school gives the schools' value added estimates. Use of the median method avoids using a regression model, which is less obvious for a non-statistical audience. Also the use of medians makes the models robust to the effect of outliers (Ray, 2006). In this median method, skewness in prior attainment or other control variables are mentioned as a problem, because the medians can be less reliable established for the more extreme levels of prior attainment. Ceiling effects for the most able pupils and the relative unstable results for small schools are other disadvantages of this model (Tymms et al., 2004).

An alternative for the median method is the use of a single level regression analysis to predict the expected value of the final achievement of children. By means of an ordinary least squares (OLS) regression analysis an expected value of final achievement can be calculated for students given other characteristics at intake. Webster et al. (1998) call a single-level regression analysis a significant improve over unadjusted raw test scores. The procedure in which the schools' value added is calculated out of the expected and achieved scores is similar as in the median method. For each child the difference between the expected and achieved score is calculated. The value added of a school can be found by averaging the deviations of the children in the school. The advantage of the regression approach above the median method is that the regression approach gives the opportunity to include other background variables on the student level more easily. A disadvantage of using regression bases value added models is the transparency of the results for a non-statistical audience.

Value added estimates based on multilevel analysis take the hierarchical structure of the data into account, where students are nested within classes and classes are nested within schools. Final achievement of students is used as the dependent variable and prior achievement and other background characteristics serve as covariates in multilevel regression models in the traditional estimations of value added. Although, depending on the context and the educational system other dependent variables and other statistical models are used. Besides the hierarchical structure of the data multilevel analysis gives the opportunity to find the part of the variance in final achievement of the children or gain scores which is associated with the school level (Creemers et al., 2003). In these models it is assumed that after controlling for several covariates the residuals on the school level represent the schools' effect.

Aitkin and Longford (1986) describe the following requirements of the analysis of school effects or value added: "The minimum requirement of an adequate analysis of school effects are: 1) pupil-level data on outcome, intake and relevant background variables, together with relevant school- and LEA (district) variables. 2) Explicit modelling of the multilevel structure through variance components at each sampling level. 3) A careful analysis of interactions between explanatory variables at different levels, of random variation among schools in the regression coefficient of pupil level variables." (p.25). Goldstein (1997) adds that studies should be longitudinal and data collected for at least three data-collection periods. Individual children in a class or a school share common experiences, which may lead to more homogeneous results than it would be in the case of a random sample of children (Aitkin et al., 1986). Therefore, using methods that do not allow to take the hierarchical structure of the data into account will overestimate the size of the school effects (Goldstein, 1997). "Analyses using such methods pose several troublesome threats to statistical conclusion validity including: aggregation bias, undetected heterogeneity of regression among sub-units, wrongly estimated parameters and their standard errors, and related problems associated with the failure to satisfy the assumptions of independence required by single-level models." (Hill et al., 1996, p. 2).

In the study of Webster et al. (1996) very strong correlations were found (reading, r=.98; mathematics, r=.96) between value added estimates derived from single-level regression analysis and multi-level analysis (Webster, Mendro, Orsak, & Weerasinghe, 1996). In a replication of this study in 1998 correlations were found around .95 between the results of single-level OLS regression analysis and multilevel analysis (Webster et al., 1998). Similar results are found in primary education in Maryland, US (Yen, Schafer, & Rahman, 1999). Very strong correlations are found between the results of single-level regression analysis and the multilevel analysis (correlations between .93 and .95 for OLS and HLM, and between .91 and .93 for WLS and HLM). These correlations depend on the number of students within schools and the amount of between school variation. Therefore, comparisons from other contexts or countries might lead to different results. These correlations between the results from different statistical models seem very high. However, in educational accountability small differences between models can have important consequences for individual schools. As a result, if biases in the estimated value added of schools can be prevented by using multilevel modelling it is very important to do so.

1.6 Validity and reliability of value added

Although many studies claim value added to be the best method we have today for the estimation of the effectiveness of schools, many studies also raise questions about the validity and the reliability of this measure. Problems with the reliability of value added led to the statement that value added can, at best, be used as a crude screening device to identify outliers, but not as a definitive statement of school effects (Goldstein, 1997). The validity of value added estimates of schools has become a very popular object of research since these models are increasingly used in high stakes educational accountability systems (McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Cantrell, Fullerton, Kane, & Staiger, 2007; Koretz, 2008; Kane & Staiger, 2008; Rothstein, 2008).

In the first place, the validity of the dependent variables, in most cases cognitive achievement of students on core subjects or standardized tests (Sharp, 2006), should be assessed. Ideally, the set of outcome measures in value added should cover the skills that are demanded by society (Meyer, 1997) and measures that reflect the educational goals of the schools (Hill et al., 1996; Coe & Fitz-Gibbon, 1998). Besides these restricted measures of cognitive skills, schools also pursue the development of personal, affective and social skills in their students (Peschar & Van der Wal, 2001). Indicators based on measurements of literacy and numeracy do not cover the total curriculum of schools, which means that the face validity of indicators based on these restricted measurements is limited.

Closely related to the validity of the dependent variable is the issue of consistency of school effects over multiple possible outcomes of education. Average examination grades are regularly used as dependent variable, although most evidence of prior research indicates a moderate level of consistency for different subject areas (Mandeville & Anderson, 1987; Mandeville, 1988; Sammons et al., 1993; Bosker et al., 1997; Thomas, Sammons, Mortimore, & Smees, 1997b; Luyten, 1998; Bosker & Luyten, 2000; Teddlie, Reynolds, & Sammons, 2000b; Ma, 2001; Luyten, 2003). "This means that when a school is successful with respect to mathematics this is not necessarily true for language, but it also implies that, generally speaking, good results for mathematics to some extent coincide with good results for language" (Luyten, 2008, p. 285). Using a single value added indicator masks the differences in effectiveness between subjects within schools (Luyten, 2003). A measure of value added for subjects or related groups of subjects will give a more detailed picture of the schools' effectiveness and will do more justice to the complex nature of schooling.

The frameworks of educational accountability systems differ in the extent to which they take account of the multiple possible outcomes of education. The Dutch framework for primary education, for example, describes indicators for average test scores and separate scores for language and mathematics (Inspectie van het Onderwijs, 2009; Inspectie van het Onderwijs, 2011b). With respect to social outcomes, only the processes within schools are assessed during school inspections. No performance indicators have been described for other than cognitive outcomes and efficiency currently. The accountability framework for Dutch secondary education only describes general indicators for grades on the final examination, although results are published for schools also on separate subjects. No specific indicators are described for non-cognitive outcomes of education. The inspection framework in the United Kingdom describes a couple of outcome indicators based on average test scores and several indicators concerning non-cognitive outcomes, such as students behavior, healthy lifestyles, enjoyment in learning, moral, social and cultural development (Ofsted, 2010).

Moreover, one could question which control variables should be taken into account. Finding the unique contribution of the school out of the other factors implicates that this school effect can be isolated statistically. If value added scores are used for comparing schools on their performance in educational accountability, estimates of school effects should take account for pupils prior achievement, family background and school composition to isolate the school effect (Willms, 1992). A Dutch exploration raises concerns about not knowing whether all relevant variables are included (Verhelst et al., 2003). The assumption of value added indicators is that after correction for prior achievement and relevant background characteristics the estimated effects of schools cannot be attributed to these control variables. However, it remains unclear if there are any other variables which might have an important impact on the estimated value added or if some of the included variables related to school practices removed some of the school effects, due to selection processes. Value added estimates should therefore be handled with great care (Verhelst et al., 2003). Several attempts have been made to develop strategies to choose appropriate background characteristics. According to Salganik (1994) the choice of background factors for the analysis of the value added estimates should be based on three criteria. First, the background factors must be related to student performance. Second, the background factors must be beyond the control of the school. And finally, the factors must be accepted by the public, educators and policymakers as legitimately related with educational challenges of schools (Salganik, 1994). However, by including control variables such as ethnicity and gender in the statistical analysis of value added,

although related to student performance, one implicitly accepts differences in achievement between boys and girls or students with different ethnic backgrounds.

Furthermore, a single value added estimate for all subgroups of students within schools might mask differential school effectiveness. Schools can be differential effective for some sub-populations of students within the school (Nuttall, Goldstein, Prosser, & Rasbash, 1989; Sammons et al., 1993; Thomas, Sammons, Mortimore, & Smees, 1997a; Veenstra, 1999; Gray, Peng, Steward, & Thomas, 2004), for example subgroups based on prior achievement, gender or ethnic background. Value added estimates for different sub-groups tend to correlate strongly, but not perfect (Thomas et al., 1997a). These strong positive correlations give an indication that in the more effective schools all students tend to perform well, but that some sub-groups benefit more (Kyriakides, 2004).

Finally, the stability of estimates of value added over successive years is regularly used as an estimate of the reliability of value added. Evidence from prior research in secondary education shows that school effects tend to be quite stable over time (Willms et al., 1989; Mortimore et al., 1994; Van der Werf & Guldemond, 1996; Thomas et al., 1997b; Teddlie et al., 2000b). Very large differences in the effects of schools between consecutive years might indicate that there are problems with the reliability of value added. Proposals for estimates of school effects measured over several years are given by many researchers (Meyer, 1997; Thomas et al., 1997a; Teddlie et al., 2000b; Inspectie van het Onderwijs, 2003; Wijnstra et al., 2003; Van de Grift, 2009).

The last remark on the validity of value added estimates is not a characteristic of value added itself, but a consequence of the use of these kinds of quality indicators in accountability systems. "If performance indicators are to be useful they must reflect the qualities that administrators and teachers want to influence, and be susceptible to improvement through changes in policies and practice" (Willms, 1992, p. 85). Willms refers to this as the intrinsic validity of school effects. The strength of the intrinsic validity is related to the possible ways that schools have to raise the scores on the indicator. Schools can raise their value added scores by providing better quality education. But there are several opportunities to raise their value added scores by gaming the system. Examples of gaming the system are reshaping the test pool, for example by increasing placement of students in special education (Figlio & Getzler, 2002; Jacob, 2005; Cullen & Reback, 2006; Lemke, Hoerandner, & McMahon, 2006; Swanborn & De Wolf, 2008), teaching to the test (Jacob, 2005) and test manipulation (Jacob & Levitt, 2003; Jacob, 2005).

1.7 Methodological challenges in modelling value added

Besides the challenges concerning reliability and validity of value added, some other methodological challenges arise. Data requirements, imperfect hierarchical structured data, ceiling effects, greater variability of the school effects for small schools and measurement errors are examples of these challenges.

The estimation of value added scores of schools requires longitudinal data on student level (Willms, 1992; Goldstein, 1997; Teddlie et al., 2000a; Bosker et al., 2001; Ray, 2006; Amrein-Beardsley, 2008). At least data on the prior achievement and final achievement for an appropriate school period are necessary for calculating the estimates of value added (Bosker et al., 2001; Roeleveld, 2003a). Limited and missing data of students can cause bias in the estimates of value added scores of schools (Meyer, 1997). Mobility of students, excluding students from making tests used for the analysis of value added, sickness at one of the test occasions are examples which can lead to missing data. Because the least able candidates are the most likely to be excluded due to missing data (Rubin, Stuart, & Zanutto, 2004), a method which includes only the complete cases will give an upward bias in the estimates of school and pupil performance (Thomas et al., 1997b).

Most of the existing value added models treat children as belonging to the school where they made their final test (Tymms et al., 2004; Goldstein, Burgess, & McConnell, 2007; Leckie, 2008). The effects of former schools on the academic achievement of these students are ignored and the final school gets all the credit or the full blame. According to Goldstein et al. (2003) the assignment of students to a single school in case of mobility can distort inferences about the effects of schools. Phenomena like student mobility and long term effects of primary schools lead to deviations from the strict hierarchical structure of data. Multiple membership models have shown that traditional value added models, in which student mobility is ignored, underestimate the effect of schools (Goldstein et al., 2007). Because of student mobility some researchers believe that the period between prior and final achievement used in value added analyses should be cut into smaller sections to minimize missing data (Meyer, 1997; Onderwijsraad, 2003; Inspectie van het Onderwijs, 2003; Wijnstra et al., 2003; Van de Grift, 2009). Schools for primary education can have a long term effect on children in secondary education (Goldstein & Sammons, 1997; Goldstein et al., 2007). Not only adjustments of prior achievement but also of all previous education should be made to find a better estimation of the short- and long-term school effects (Kyriakides & Creemers, 2008).

Ceiling effects are a major problem in the estimation of value added scores for schools (Schagen, 2006). In case of ceiling effects students with high prior achievement are not likely to reach scores far above the predicted scores but are likely to get results much more below their expected scores. Bias in the estimates of value added can be of specific importance for school types which serve an atypical population of students at both ends of the distribution. This can result in an underestimation of the value added estimates of schools with a relative large number of high prior achievement students (OECD, 2008) and an overestimation of the value added score for schools with a relative large number of low prior attainment students.

Greater variability and instability of school effects of very small schools can pose problems for the use of these school effectiveness measures in educational accountability systems. Small schools have a greater chance of getting very high or very low value added estimates. In a process called shrinkage, the residuals of schools in multilevel models, or the schools estimates of value added, are shrunken towards the mean (Goldstein, 1997; Tate, 2004). Population information is used in this process in addition to the group or school information (Snijders & Bosker, 1999). The residual of a school is therefore pulled a bit towards the general mean. The proportion of the shrinkage depends on the reliability of the estimated school effect (Hox, 2002). The reliability of the estimated school effect depends on the group size of the school and the distance between the overall group estimate and the school-based estimate of the coefficients. The shrinkage will be greater for schools with a small sample size. The residuals for large schools will be almost the same as the group mean. Shrinkage will be greater as the difference between the population estimates and the school estimate is larger. In other words, shrinkage is larger for the most and least effective schools.

Furthermore, most of the background factors cannot be measured directly and proxies are used instead. For example, in the UK, social class is commonly measured by the entitlement to free school lunches, which cannot be considered as an accurate proxy for social class or income (Ray, 2006). This is an example of underspecification of this background characteristic. Aggregated variables at the school level suffer from measurement error as well, namely sampling error (Woodhouse, Yang, Goldstein, & Rasbash, 1996). The assumption of control variables in regression analysis is that they are measured error free. The use of unreliable explanatory variables produces a bias towards zero in the estimation of the regression coefficients for these variables, and corresponding biases in the coefficients of other variables in the model (Aitkin et al., 1986; OECD, 2008). "Estimates of school-, class- and student-level will be influenced by underspecification of (i.e. not measuring all relevant aspects of family background characteristics) and unreliability in measures of intake factors (resulting in attenuation

of regression coefficients). The impact of not measuring all relevant intake characteristics, by definition, is unknown, but may be assumed to result in biased estimates. […] The impact of unreliability in intake measures invariably leads to over-estimates of the proportion of variance at the student-level and to under-estimates of the effects at higher levels." (Hill, et al., 1996, p. 10). Estimates and conclusions can vary depending on the amount of measurement error in the explanatory variables at level 1 and level 2 and in the response variable (Woodhouse et al., 1996; Goldstein, Kounali, & Robinson, 2008).

1.8 The present dissertation

The present dissertation examines the usefulness of value added as an indicator of the performance of schools in educational accountability for valid comparisons between schools and possibilities for risk based accountability. The studies in this dissertation focus on primary, secondary and senior secondary vocational education in the Netherlands. In each study, a specific aspect of value added will be studied in more detail to investigate to what extent value added is useful for educational accountability and which weaknesses should be acknowledged.

The study presented in Chapter 2 explores the differences in the conceptual meaning and empirical estimates of value added for particular sets of control variables, also known as types of school effects. In this study the following research question is addressed: What is the impact of the choice of control variables for differences in intake of students on the estimations of value added? Longitudinal data from the Cohort Studies in Secondary Education (VOCL'99) were used to estimate value added with multiple sets of control variables. Differences in classification of the various value added models in terms of underperforming, average and overperforming are used to illustrate the importance of control variables.

Chapter 3 attempts to shed light on the impact of imperfect hierarchical structures of data in education on the estimated value added of secondary schools. In this chapter, the following research question is investigated: What is the effect of more realistic modeling of the hierarchical structures in education on the estimation of value added? Two phenomena, student mobility and long term effects of primary schools, which cause imperfect hierarchies in educational data were modeled using subsets of data from the Cohort studies in Secondary Education (VOCL'99). The association between the estimated value added from different models give an indication of the

impact of these phenomena on the validity of traditional estimation methods of value added.

In Chapter 4 a study is presented in which value added on cognitive and non-cognitive outcomes of education in secondary education were compared. The aim of this study was to develop value added indicators for several cognitive and non-cognitive outcomes and to test the association between the estimated value added of schools for those outcomes. Value added and educational effectiveness research (EER) are widely criticized for the narrow focus on cognitive outcomes, language in the mother tongue and mathematics in particular. Value added estimates were calculated for multiple outcomes (affective, social and cognitive). Multivariate multilevel analysis showed both the variation in value added for the multiple outcomes and their association.

Value added is usually based on the performance of tests or examination. However, test based performance indicators are widely criticized for the possibilities for strategic behaviour of teacher and schools to artificial inflate test scores. In the fifth chapter differences in the effects of schools on the educational positions of their students as measured through "leerjarenladder" are investigated. This value added indicator based on the educational positions of students leads to opposite incentives for schools with respect to strategic behaviour. Combinations of multiple indicators, both for test performance and the educational position, are suggested to increase the robustness for strategic behaviour of schools.

Chapter 6 presents an explorative study on the development of value added in the vocational education sector. The aim of this study was to develop a value added indicator for vocational education based on a multilevel model. A subsample was used from the national pupil database (BRON) to investigate differences in value added between educational institutions and clusters of training programmes. Multinomial logistic multilevel regression analysis was applied to deal with the level of diploma as outcome measure.

Recently, risk based approaches of educational accountability are implemented in the Netherlands. A risk based strategy implies that school inspections can vary across schools depending upon the results of previous inspections and their subsequent performance. The seventh chapter of the dissertation examines whether future "at risk" schools in primary education can be accurately identified based on all current information of schools, their staff and student population. Data from a Monitoring and Evaluation system is used to estimate value added for a sample of primary schools.

Finally, Chapter 8 summarizes the main findings of the studies presented in this dissertation. Furthermore, some general conclusions, limitations of the studies, suggestions for future research and practice are discussed here.

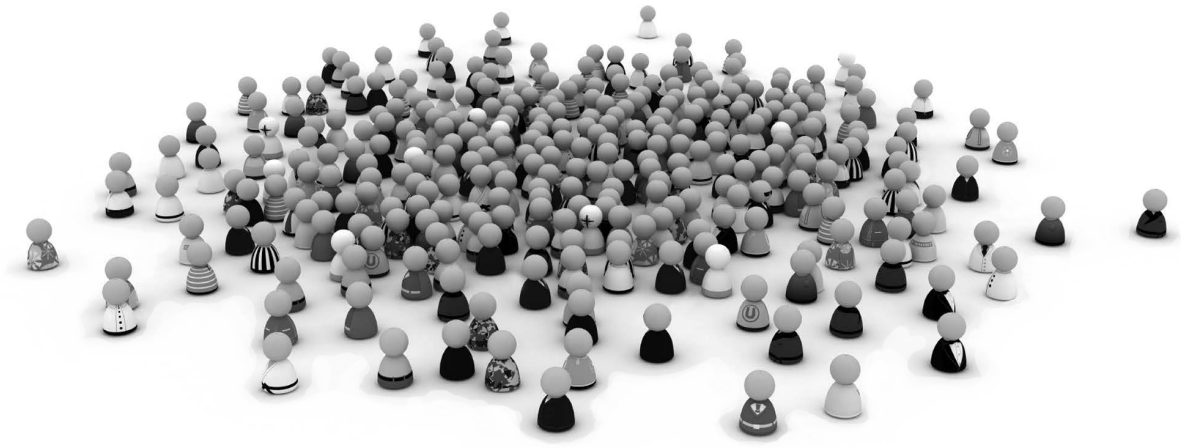
1.8.1 A note to the reader

The empirical chapters in this dissertation are written in such a way that they can be read independently. Consequently, some overlap in the introductory sections of the

Chapter 2

Conceptual and Empirical Differences among Various Value Added Models for Educational Accountability

This chapter is published as:
Timmermans, A.C., Doolaard, S. & De Wolf, I. (2011). Conceptual and empirical differences among various value added models. *School Effectiveness and School Improvement, 22*, 393 – 413. DOI: 10.1080/09243453.2011.590704



Abstract

Accountability systems in education generally include indicators of student performance. However, these indicators often differ considerably among the various systems. More and more countries try to include value added measures, mainly because they do not want to hold schools accountable for differences in their initial intake of students. This study presents a conceptual framework of these value added measures, resulting in an overview of five different types. Using data from Dutch secondary schools, we empirically provide estimates of these different measures. Our analyses show that the correlation between the different types of school effects estimated is rather high, but that the different models implicate different results for the individual schools. Based on theoretical considerations, arguments are given to use the following indicators in the value added accountability models: prior achievement, student level background characteristics, and compositional characteristics of the student population.

2.1 Introduction

During the last decade, many countries have introduced systems of school accountability by which schools have to give account of their educational practices and policies to authorities. In several countries the focus of these educational accountability systems is on the schools' output. Generally, student achievement is used as the output measure. This situation has generated the interest of scientists as well as politicians in the development of output indicators of school effectiveness to estimate the quality of schools. Many of the proposed output indicators try to isolate a school's effect on students' learning or progress. This means that the schools can only be held accountable for their effects on these issues and not for the effects of factors that are outside their control, such as the students' prior education or their family's socio-economic status. Methods that account for the differences in student intake among schools are usually called value added. For authorities the aim of value added indicators in an accountability system is to identify possible weak schools. Value added models differ in the variables used to control for differences in student intake among schools. Techniques to estimate value added have been developed in an ongoing process since the 1980s (Raudenbush et al., 1986; Aitkin et al., 1986; Willms et al., 1989; Hill et al., 1996; Meyer, 1997; Goldstein, 1997). In the educational accountability systems in some countries or states, value added estimates of school effectiveness are already being used. Well-known examples are Contextualized Value Added (CVA) (Ray, 2006), used in the United Kingdom, and the Tennessee Value Added Assessment System (TVAAS) (Sanders & Horn, 1994), used in several states of the United States.

The different types of value added models might produce different results for the individual schools. Especially in countries where these indicators are used or developed for assessing schools' performance output, knowledge about these differences is important in choosing the best suitable model. Furthermore, it might be important for researchers in school effectiveness research to choose the type of value added model that fits with specific research questions.

2.1.1 Research questions

The aim of this study is to compare the different operationalizations of the value added measures from both a theoretical and an empirical perspective. We have tried to answer the following research questions:

1. What are the theoretical or conceptual differences among the different value added models?
2. What are the empirical differences among the different value added models?
3. Is there a 'golden standard' for the value added models used in accountability systems?

For the empirical part of the study (question 2), performance data from Dutch secondary schools were used.

2.1.2 Value added models and the use of covariates to control for differences in intake among schools

The various value added models are all based on the notion that student achievement is influenced by a number of factors, namely students' background characteristics, school context, school practices, and students' unique contribution (Keeves, Hungi, & Afrassa, 2005). In other words, a model for estimating school effects should recognize that the learning process of children is a function of the exposure to multiple social contexts (Downey, Von Hippel, & Hughes, 2008). In general, a value added model accounts for differences in intake among schools with respect to valuing their outputs: their students' achievements. In some models, adjustments are made only for prior achievement, while other models also include various types of student and school characteristics. Because value added models differ in the control variables used to estimate the value added, they differ in the meaning of their output: a residual that indicates how well a school is doing compared to other schools. Based on literature five different value added models are described below. Table 2.1 presents an overview of these models. The first model, which depicts the gross school effect, is a model without control variables. We define this basic model as a 'type 0' model. It gives an indication of the difference in the average performance of students between school J and the average performing school, regardless of the differences in intake between the schools. Strictly, this model cannot be called value added. A 'type 0' school effect can be estimated using an empty multilevel model, which estimates the unadjusted averages of the individual performance of students at a school. However, it does contain learning, which is - due to many factors - beyond the control of the school

(Meyer, 1997; Webster et al., 1998), while the influence of these factors is unevenly distributed among the schools (Hill et al., 1996). Therefore, the 'type 0' value added model does not seem appropriate to be used by authorities for accountability purposes, as it does not provide a fair comparison among the schools.

The second model can be defined as a type AA value added model. This model controls for prior achievement. In doing so, this type measures the difference between school J and the average school for students with a comparable level of prior achievement. Much of the association between potential confounders and the final outcome of schooling can be removed by using a reliable pre-test of achievement, as shown by Raudenbush (2005). Using only prior achievement as a control variable, the value added in this approach implicitly refers to learning gains during secondary education. Whether or not the effect is solely caused by the school is another matter. The Tennessee Value Added Assessment System is an example of a value added indicator that only includes prior achievement in assessing the effectiveness of teachers. We define this factor as a 'type AA' school effect.

Table 2.1
Overview of value added models

Type of value added models	Control variables	Meaning of the school effect
Type 0	-	The difference between school J and the average school in the average performance of students; this is equivalent to the gross school effect
Type AA	Only prior achievement (or aptitude-related covariates)	The difference between school J and the average school for students with a certain level of prior achievement
Type A	Only student level covariates, in any case prior achievement and some background characteristics associated with student achievement	The difference between school J and the average school for children with certain characteristics and similar levels of prior achievement
Type B	Student level covariates and compositional covariates of the student body on the school level	The difference between school J and the average school while controlling for student characteristics and with a similar context
Type X	Student level covariates, compositional covariates on the school level and other non- malleable school level covariates	The difference between school J and the average school while controlling for student characteristics, with a similar context and similar non-malleable characteristics

A third model is that of Raudenbush and Willms (1989), referred to as the 'type A' school effects model. In this value added model adjustments are made for student level covariates, such as students' background characteristics and prior achievement (Willms et al., 1989). According to Willms (1992), in order to isolate the school effect, prior achievement and family background should in any case be used as control variables. The 'type A' school effect model answers the question what the difference is between school J and the average school for children with comparable characteristics. However, 'type A' is not only influenced by the quality of the teaching staff and educational practices, but can also be impacted by the more positive or negative surroundings of the school or the schools' population. Willms and Raudenbush therefore argue that it would not be fair to compare schools on the basis of type A school effects because the quality of teaching is only a part of this school effect. Furthermore, there might be another problem with the 'type A' and 'AA' models. The use of student level covariates might cause problems when selection effects occur. If in a non-experimental setting the assignment to groups or schools is related to the potential benefit of being assigned to these groups or schools, statistical controls for measured confounders are not sufficient in removing bias (Raudenbush, 2005). The fact that in the case of the 'type A' school effect only student level covariates are used to control for initial differences in intake might lead to distortions in the estimates if there is a relation between the student level covariates and the causal effects of the school. For example, if students from advantaged families (high socio-economic status) tend to choose more effective schools, using this covariate in the model will also explain the part which should be a school effect.

To control for effects of compositional covariates of the student body, often a fourth value added method is used. Raudenbush & Willms (1989) refer to this model as the 'type B' value added model. This 'type B' model controls for student level characteristics and compositional covariates. Conceptually, the effects of the specific practice and policy of a school on a child's performance are isolated in the type B effect. This school effect relates to the difference between school J and the average school while controlling for student characteristics and a similar student composition (Raudenbush & Willms, 1995). In the estimation of the 'type B' school effect adjustments are made for student and school level covariates (Willms et al., 1989). Examples of school level covariates are the average prior achievement and socio-economic status of the students as well as any possibly wider social influences, also called school composition. According to Willms (1992), because of the isolation of school practices, the 'type B' effect is the effect mostly preferred in comparing schools by means of educational accountability systems. The Contextualized Value Added

model from the UK is an example of a model used in estimating the type B school effect. Despite using covariates of student characteristics and school composition we can, however, never be sure whether a school with a high estimated school effect is really an excellent school. There can still be other explanations for these effects, for example professional homework support outside the school.

Besides the selection effects covered in the 'type B' value added model, there might also be other covariates related to the educational practices within schools which explain part of the school effects. Teaching practices depend heavily on the student body within the class. They can vary considerably between student groups performing at a high and a low level. Similarly, teaching can vary a great deal between hetero- and homogeneous groups of students. Controlling for school composition in the type B effect might remove some of the joint effects of school composition and school policy and practices (Willms, 1992; Opdenakker et al., 2001). Research has shown that although school composition largely and uniquely contributes to student outcomes, part of the effects of school composition are mediated by practices (Cervini, 2005; Opdenakker & Van Damme, 2007). Controlling for school composition might lead to a further underestimation of the true school effects. On the other hand, estimating school effects without school composition as covariate might lead to an overestimation of these effects because of the unique contribution of school composition, as students - for example - may stimulate one another.

The fifth and last value added model is the 'type X' model. The type X effect is a further specification of the 'type B' effect. The 'type B' effect tries to isolate the school practices, while the 'type X' effect only isolates the effects of the school practices on which the school has influence. This means that adjustments are made for non-malleable factors over which the school has no control, such as school size, as well as urban or rural and stratifying factors, such as state or school type. Student mobility, overcrowding, and staffing patterns are also suggested as covariates for value added indicators of school effectiveness (Webster et al., 1996). Keeves et al. (2005) state that the 'type X' effect is the most appropriate estimate of school effects to be tested by an educational accountability system. For the 'type X' effect similar problems with the use of covariates arise as for the 'type B' effect. If a value adding indicator is used in an educational accountability system it is questionable whether non-malleable covariates should be used. The aim of such an indicator is to identify possible weak schools, regardless of the location within the country or other non-malleable factors proposed, which is why it is doubtful whether this type of school effect should be used by authorities in their educational accountability system.

In summary, we distinguish between five types of value added models, ranked by increasing complexity. All models use a multilevel approach to estimate school effects, but differ in their types of control variables, namely a basic model without control variables (type 0), a model which controls for prior achievement (type AA), for prior achievement and other student level covariates (type A), for student level covariates and compositional covariates (type B) and for student level covariates, compositional covariates, and other non-malleable school level covariates (type X). Possible advantages and disadvantages of these models have been identified. In the forthcoming section the empirical differences among the five value added models are compared to establish which one is the most suitable for identifying underperforming schools.

2.2 Methods

2.2.1 Subjects

The data used in this empirical study were collected as part of a national longitudinal study in secondary education in the Netherlands, the "Cohort Studies in Secondary Education" (Dutch abbreviation: VOCL). The data concerned students who entered the first grade of Dutch secondary education in the Netherlands (comparable to the 7th grade in the United States) in the year 1999, also called the VOCL'99 cohort. The total cohort consists of a sample of approximately 20,000 students. This sample has been considered as representative of the schools and students in the Dutch secondary education (Kuyper & Van der Werf, 2003b).

We performed separate analyses for two school types: the prevocational education theoretical track (VMBO tl) and higher general secondary education (HAVO). The Dutch secondary education system consists of multiple differentiated tracks, for which the students are selected at age twelve on the basis of their scholastic aptitude. VMBO tl is one of the four year vocational tracks preparing students for vocational education. The HAVO track is five years long and prepares students for professional education.

For the current study we selected a subsample from the VOCL'99 population based on the following criteria: for each student an identification variable at both the student and the school level had to be available as well as a central examination score, and he/she had to belong to a school represented by 10 or more students in the sample. For both school types 90% of the total sample was included in the analysis.

Table 2.2 gives an overview of the number of students and schools in the sample for the separate school types. In addition to the data from the VOCL'99 cohort, a dataset of the Dutch Inspectorate of Education was used for the school level variables as proposed in the type X effects model. These school level variables were merged with the VOCL'99 database by the school level identification variable.

Table 2.2
Overview of samples for school types used in further analyses

	VMBO tl	HAVO
Number of students	3868	3428
Number of schools	77	60

2.2.2 Variables

Student level characteristics The data in the VOCL'99 cohort were derived from several sources and on several occasions (Kuyper & Van der Werf, 2003a). The overall mean grade on the central exam was used for the study of the differences among the five different types of school effects (0, AA, A, B and X). This variable served as the dependent variable in the multilevel regression analysis.

Halfway through the seventh grade the "cito-entree" test took place. The cito-entree test has been developed by CITO, the Netherlands Institute for Educational Measurement. This test contains the parts Dutch language, mathematics, and information processing. For our study, the total score on the "cito-entree" test was used as an overall measure of prior achievement. The total test had a reliability (Chronbach's α) of .90 (Kuyper et al., 2003b) and the range of possible scores on this test lay between 10 and 60 points. Information on the student's intelligence was gathered by the "Groninger Intelligence test for Secondary Education" (Dutch abbreviation: GIVO) (Van Dijk, 1995). This test is administered in the eighth grade. The GIVO intelligence test consists of verbal and symbolic sections. A reliability (Chronbach's α) of .91 was reported for the verbal part of the intelligence test and .93 for the symbolic part for a sample of students in secondary education covering all school tracks (Evers, Van Vliet-Mulder, & Groot, 2000). Socio-economic status was measured by the highest educational level completed by one or both of the student's parents. This variable consisted of six categories, ranging from only primary education to post-graduate. In the analysis this item was used as a continuous variable.

Furthermore, the age of the students measured in years was used as a covariate. Tables 2.3 and 2.4 present the descriptive statistics for these variables.

A dummy was created for the student characteristic gender, in which boys (49.5% in the VMBO sample and 46.4% in the HAVO sample) formed the reference group. For the dummy variable ethnicity native students served as the reference group. In the VMBO sample 83.5% of the students were native and in the HAVO sample 84.7%. For second language two dummy variables were created, with only Dutch speaking students as the reference group (78.4% for VMBO and 81.7% for HAVO). Other categories for this variable were 'bilingual' (6.5% and 5.6%) and 'only another language or dialect' (15.0% and 12.7%).

Table 2.3
Descriptive statistics of variables used in the analyses of school effects VMBO (tl)

	Mean	Standard deviation	Minimum	Maximum
Final achievement	6.40	0.67	3.08	8.70
Prior achievement	35.55	7.30	10.0	60.0
Intelligence	101.12	9.06	65.0	140
SES	4.01	0.993	2.0	7.0
Age	12.99	0.44	11.6	15.4
Prior achievement (school average)	35.22	2.90	25.75	40.30
Prior achievement (school standard deviation)	6.90	0.93	5.28	9.11
IQ (school average)	100.77	5.01	88.33	113.00
IQ (school standard deviation)	8.15	1.80	4.24	13.43
SES (school average)	3.97	0.34	2.61	4.63
SES (school standard deviation)	0.96	0.17	0.52	1.51
Type of neighbourhood	2.47	1.57	1.00	7.74

School level characteristics Variables such as the average prior achievement of students per school type, the average socio-economic status, the average intelligence, and the number of students were created by aggregating the student data to the school level. Also the mean type of neighbourhood where the students lived was used as a compositional variable on the school level. This variable was coded in the opposite direction with a range from 1 (normal neighbourhood) to 8 (accumulating to problem neighbourhood). This variable was based on the proportion of inhabitants from non-

Western origin, the proportion of inhabitants living from social services, and the proportion of inhabitants with a low income. In addition to the averages, the standard deviations of background characteristics on the school level were used to assess the possible effects of homogeneous versus heterogeneous school populations. Denomination was a dummy for different religious beliefs, with public schools as reference group. School structure was a dummy in which the schools with the least amount of possible school tracks were used as the reference group. This variable gave an indication of whether the analyzed school type was the only school type taught at the specific school or that multiple types of school tracks could be attended.

Table 2.4
Descriptive statistics of variables used in the analyses of school effects HAVO

	Mean	Standard deviation	Minimum	Maximum
Final achievement	6.27	0.71	3.64	9.13
Prior achievement	42.77	6.37	14.0	59.0
Intelligence	109.12	9.57	80.0	143.0
SES	4.35	1.02	2.0	7.0
Age	12.89	0.38	11.4	15.5
Prior achievement (school average)	42.65	2.27	36.28	47.75
Prior achievement (school standard deviation)	5.88	0.94	3.86	7.77
IQ (school average)	108.98	4.28	101.60	125.00
IQ (school standard deviation)	8.92	1.39	6.60	14.80
SES (school average)	4.32	0.35	3.14	5.00
SES (school standard deviation)	0.95	0.22	0.00	1.43
Type of neighbourhood	2.43	1.65	1.00	8.00

2.2.3 Method of analysis

We used hierarchical linear models for estimating school effects using MLwiN software (Rasbash, Steele, Browne, & Goldstein, 2009). These models are considered the most appropriate because they take the hierarchical structure of the data into account (Snijders et al., 1999). Restricted Maximum Likelihood (ReML) methods were used to estimate the parameters. First we estimated an empty model for the gross school effect (type 0). For analyzing the types of value added models we then estimated a two-level model by first fitting the student level variables, which led to an

estimation of the type AA and type A effects, and then fitting the school level variables, which resulted in the type B and type X effects. In the VOCL'99 cohort some data were missing for the explanatory variables. This is why some of the extended models were based on fewer cases, which made it impossible to compare indices of model fit. In the analysis the continuous variables were centred round their grand mean. We used a two-level model in which the students (level 1) were nested within the schools (level 2), although we recognized that the true hierarchical structure of education data is much more complicated. Next, the posterior residuals (the estimated difference between the average school and other schools) from these analyses were saved and imported into SPSS for further analysis. For comparing the different types of school effects correlations and Kappa's were calculated to estimate the strength of the association.

Table 2.5

Multilevel regression analysis for the estimation of different types of school effects in the pre vocational theoretical track (VMBO tl)

	Gross school effect (type 0)	Type AA	Type A
Student level fixed effects			
Constant	6.391 *(0.030)**	6.401 *(0.024)**	6.433 *(0.030)**
Prior achievement		0.035 *(0.001)**	0.023 *(0.002)**
Other student background variables			
Intelligence			0.022 *(0.002)**
Age			-0.073 *(0.029)**
Socioeconomic status			0.069 *(0.012)**
Ethnicity (reference group native students)			0.058 *(0.039)*
Gender (reference group boys)			0.011 *(0.024)*
Second language (reference only Dutch speaking students)			
Biligual			-0.086 *(0.051)*
Only other language			-0.003 *(0.086)*
Only dialect			-0.050 *(0.045)*
Random effects			
School level variance	0.056 *(0.011)*	0.035 *(0.007)*	0.031 *(0.008)*
Student level variance	0.402 *(0.009)*	0.346 *(0.008)*	0.303 *(0.009)*
Model fit			
-2loglikelihood	7602.483	6654.779	3941.147
Number of cases	3868	3675	2342

* $p<0.05$ (two-tailed); (av) = school level average, (sd) = standard deviation at the school level
Unstandardized coefficients are reported.

Table 2.5 (Continued).
Multilevel regression analysis for the estimation of different types of school effects in the pre vocational theoretical track (VMBO tl)

	Type B	Type X
Student level fixed effects		
Constant	6.434 *(0.028)**	6.499 *(0.113)**
Prior achievement	0.023 *(0.002)**	0.023 *(0.002)**
Other student background variables		
Intelligence	0.022 *(0.002)**	0.022 *(0.002)**
Age	-0.071 *(0.029)**	-0.070 *(0.029)**
Socioeconomic status	0.065 *(0.012)**	0.064 *(0.012)**
Ethnicity (reference group native students)	0.063 *(0.040)*	0.061 *(0.040)*
Gender (reference group boys)	0.010 *(0.024)*	0.007 *(0.024)*
Second language (reference only Dutch speaking students)		
Biligual	-0.080 *(0.051)*	-0.082 *(0.051)*
Only other language	0.011 *(0.086)*	0.009 *(0.086)*
Only dialect	-0.037 *(0.045)*	-0.041 *(0.045)*
School level fixed effects		
Compositional or context variables of the student body		
Prior achievement (av)	0.004 *(0.014)*	-0.000 *(0.015)*
Prior achievement (sd)	0.001 *(0.030)*	-0.003 *(0.031)*
Intelligence (av)	-0.005 *(0.008)*	-0.003 *(0.008)*
Intelligence (sd)	-0.014 *(0.016)*	-0.010 *(0.017)*
Socio economic status (av)	0.247 *(0.103)**	0.258 *(0.105)**
Socio economic status (sd)	0.063 *(0.176)*	0.179 *(0.182)*
Type of neighbourhood (av)	0.013 *(0.025)*	0.015 *(0.026)*
Non-malleable factors		
Number of students per school		
Structure of the school (reference group schools with only theoretical track of pre vocational education)		-0.001 *(0.001)*
All pre vocational education tracks		-0.180 *(0.152)*
Pre-university, higher general secondary education and the theoretical track of pre vocational education		-0.099 *(0.119)*
All types of school tracks		-0.103 *(0.106)*
Denomination (reference group public schools)		
Catholic		0.062 *(0.075)*
Protestant		0.103 *(0.081)*
Other		-0.032 *(0.072)*
Random effects		
School level variance	0.025 *(0.006)*	0.023 *(0.006)*
Student level variance	0.303 *(0.009)*	0.303 *(0.009)*
Model fit		
-2loglikelihood	3932.368	3911.586
Number of cases	2342	2335

2.3 Results

We performed analyses for two types of secondary education: (a) the prevocational theoretical track (VMBO tl) and (b) higher general secondary education. Table 2.5 presents the results of the analysis of the school effects for the prevocational theoretical track. An intraclass correlation of .122 was found for the gross school effect (type 0). This finding gave an indication of the dependency of the scores of the students' final achievement on the schools. About 12% of the variance in the final achievement of students can be ascribed to the schools. This means that the largest amount of variance occurs at the student level. Earlier school effectiveness studies conducted in Dutch secondary education (Luyten, 1998; Veenstra, 1999) have shown intraclass correlations of similar magnitudes.

For the 'type AA' effect 'prior achievement' was used as covariate to estimate the school effects. Students with an average prior achievement had a mean score of 6.401 on their final examination. Students with above average scores on prior achievement tended to score better on the final examination. Almost 17% of the variance in final achievement on the student level was explained through prior achievement. In addition, prior achievement also explained almost 35%[1] of the variation in final achievement among schools.

In addition to prior achievement several other background variables on the student level were used as covariates in the analysis of the type A school effect. All student level covariates together explained approximately 27% of the variation in final achievement on the student level. Furthermore, these student level covariates (background and prior achievement) explained 42% of the variance among the schools. Significant effects were found for intelligence, age, and socio-economic status. In addition to the effects of prior achievement there seemed to be a tendency that students with an above average intelligence performed better on their final examination. Further, students with parents educated above average tended to have better scores on final achievement. The effect of age showed the opposite direction. Older students appeared to perform less good on their final examination. This effect can partly be explained by the fact that the older students in the sample were students who had repeated a class in primary or secondary education and that although they had received a year more education than the other students, they still performed below average.

[1] The average number of students within the schools (50) was used to calculate the explained variance on the school level.

With respect to the type B school effect, including also compositional variables at the school level, only significant effects were found for the average socio-economic status. On top of the effects of the student level covariates, students performed better on final achievement if they were taught in schools where the parents of the students were educated above average. Other compositional variables at the school level did not have significant effects. For the type X effect, non-malleable characteristics of schools were used as covariates in the analysis. It appeared that none of these covariates had a significant relationship with the achievement measure. In the case of this sample, a value added model estimating 'type X' school effects had no additional value compared to a value added model estimating the 'type B' school effect.

Table 2.6 presents the results from the analyses of the school effects for the higher general secondary education segment. Overall, they are quite similar to those for the prevocational theoretical track (VMBO tl), although there are a few remarkable differences. The intraclass correlation shows less variance in final achievement at the school level for HAVO than for VMBO tl. For the higher general secondary education there is only 7% variance among the schools.

We used prior achievement as covariate to estimate the type AA school effects. The table shows a mean score on the final examination of 6,273 for students with an average prior achievement. Students with an above average prior achievement perform better on their final examination. Almost 9% of the variance on the student level and 15%[2] on the school level is explained by prior achievement. Similar to the prevocational theoretical track also other schools differ in their intake of students with respect to prior achievement.

[2] The explained variance on the school level was calculated based on the average number of students within schools (57).

Table 2.6

Multilevel regression analysis for different types of school effects for higher general secondary education (HAVO)

	Gross school effect (type 0)	Type AA	Type A
Student level fixed effects			
Constant	6.268 *(0.028)**	6.273 *(0.026)**	6.365 *(0.032)**
Prior achievement		0.033 *(0.002)**	0.021 *(0.003)**
Other student background variables			
Intelligence			0.018 *(0.002)**
Age			-0.074 *(0.038)**
Socioeconomic status			0.062 *(0.015)**
Ethnicity (reference group native students)			-0.128 *(0.049)**
Gender (reference group boys)			-0.054 *(0.029)*
Second language (reference only Dutch speaking students)			-0.157 *(0.063)**
Biligual			0.059 *(0.119)*
Only other language			-0.021 *(0.057)*
Only dialect			
Random effects			
School level variance	0.036 *(0.008)*	0.030 *(0.007)*	0.021 *(0.007)*
Student level variance	0.475 *(0.012)*	0.437 *(0.011)*	0.392 *(0.012)*
Model fit			
-2loglikelihood	7270.642	6709.606	4011.27
Number of cases	3428	3296	2081

* $p<0.05$; (two-tailed;) (av) = school level average, (sd) = standard deviation at the school level
Unstandardized coefficients are reported.

Table 2.6 (Continued)

Multilevel regression analysis for different types of school effects for higher general secondary education (HAVO)

	Type B	Type X
Student level fixed effects		
Constant	6.337 *(0.035)**	6.361 *(0.163)**
Prior achievement	0.021 *(0.003)**	0.021 *(0.003)**
Other student background variables		
Intelligence	0.018 *(0.002)**	0.018 *(0.002)**
Age	-0.074 *(0.039)*	-0.076 *(0.039)*
Socioeconomic status	0.060 *(0.015)**	0.060 *(0.015)**
Ethnicity (reference group native students)	-0.118 *(0.049)**	-0.119 *(0.049)**
Gender (reference group boys)	-0.055 *(0.029)*	-0.053 *(0.029)*
Second language (reference only Dutch speaking students)		
Biligual	-0.147 *(0.063)**	-0.147 *(0.063)**
Only other language	0.069 *(0.119)*	0.070 *(0.119)*
Only dialect	-0.023 *(0.059)*	-0.007 *(0.058)*
School level fixed effects		
Compositional or context variables of the student body		
Prior achievement (av)	-0.012 *(0.019)*	0.004 *(0.021)*
Prior achievement (sd)	-0.013 *(0.036)*	-0.008 *(0.036)*
Intelligence (av)	-0.015 *(0.010)*	-0.012 *(0.010)*
Intelligence (sd)	0.028 *(0.025)*	0.025 *(0.026)*
Socio economic status (av)	0.019 *(0.097)*	0.004 *(0.097)*
Socio economic status (sd)	-0.255 *(0.255)*	-0.186 *(0.211)*
Type of neighbourhood (av)	-0.052 *(0.028)*	-0.053 *(0.031)*
Non-malleable factors		
Number of students per school		-0.001 *(0.001)*
Structure of the school (reference group school with pre-university and higher general secondary education)		
All school types		0.039 *(0.151)*
Pre-university track, higher general secondary education and the theoretical track of pre vocational education		0.110 *(0.155)*
Denomination (reference group public schools)		
Catholic		-0.093 *(0.080)*
Protestant		-0.086 *(0.083)*
Other		-0.118 *(0.078)*
Random effects		
School level variance	0.021 (*0.006)*	0.014 *(0.005)*
Student level variance	0.394 *(0.012)*	0.392 *(0.012)*
Model fit		
-2loglikelihood	4000.982	3996.412
Number of cases	2081	2081

In the estimation of type A school effects the models show notable differences between prevocational education and higher general secondary education. Besides significant effects for prior achievement, intelligence, and socio-economic status, also ethnicity and bilingual students have considerable impacts. Foreign students perform on average -0.128 point lower on their final examination. Native Dutch speaking students score on average 0.157 point higher at their final examination than bilingual students. The student level covariates account for approximately 19% of the variance on the student level and for 37% of the variance on the school level.

Comparable to prevocational education tl, no significant effects can be observed for the non-malleable characteristics of schools. After controlling for other student and school level covariates, we see no significant differences among the groups of schools in mean final achievement based on school structure or denomination

Table 2.7

Correlations between estimates for different types of school effects for prevocational education (VMBO tl) and higher general secondary education (HAVO)

	VMBO tl			
	Type 0	**Type AA**	**Type A**	**Type B**
Type AA	.928***			
Type A	.724***	.859***		
Type B	.626***	.759***	.939***	
Type X	.634***	.755***	.904***	.959***
	HAVO			
	Type 0	**Type AA**	**Type A**	**Type B**
Type AA	.949***			
Type A	.756***	.816***		
Type B	.654***	.714***	.898***	
Type X	.630***	.684***	.858***	.959***

***p<0.001 (two-tailed); VMBO tl n=67; HAVO n=49

2.3.1 Comparing the different value added models

The first approach to comparing the different value added models was estimating the correlations between the school level residuals. Table 2.7 presents the correlations between the different operationalizations of the school effects. These correlations can give an indication to which extent the value added models measure a similar construct. However, due to the large uncertainty associated with the estimated effects of the

individual schools, these correlations do not show the implications of the different value added models for the accountability practices of authorities.

Both prevocational education and higher general secondary education show strong correlations between the different types of school effects. Overall, the correlations are somewhat stronger for prevocational education. This finding suggests that the different types of school effects measure the same underlying construct. The correlations between the type AA and the other types of school effects vary between .742 and .859 for prevocational education, and between .684 and .816 for higher general secondary education. Even the association between the gross school effect and the type AA school effect is very strong. Both for prevocational education and higher general secondary education the gross school effect appears to correlate less strong with the other types of school effects.

These very strong correlations between the school effects estimated might partly be an artefact of shrinkage to the mean. Shrinkage to the mean or Empirical Bayes estimates means that the residuals from the multilevel analysis are pulled towards the mean on the basis of the reliability of the coefficient estimation (Hox, 2002). Especially schools with a smaller sample of students and with more extreme posterior residuals will be pulled towards the mean.

An important drawback of value added models concerns the large confidence intervals associated with the estimated school effect. Overlapping confidence intervals is why residuals from these models cannot be used for the ranking of schools in league tables (Goldstein & Spiegelhalter, 1996; Leckie & Goldstein, 2009). For assessing the implications of using different value added models in accountability systems, at best three groups of schools can be defined based on the confidence intervals (95%) around their estimated effect: (a) average schools, (b) overperforming schools, and (c) underperforming schools. The average schools, the largest group, are schools for which the confidence intervals around their estimated school effect (residual) contain zero, which means that these schools cannot be statistically distinguished from the average. A small group of schools performs better than average. For these schools the confidence intervals around the estimated school effects are located above zero. These schools can be distinguished as effective above average, thus as overperforming. Similarly, a small group of schools can be distinguished as less effective. The confidence intervals for these schools are located below zero, which means that these schools are underperforming. With respect to accountability, this small group might be of particular interest in the identification of failing schools. Table 2.8 shows an example of the classification of schools into three

groups for HAVO for the 'type A' and 'type B' effect. The schools on the diagonal of the table are classified by both models into the same categories of schools. Outside the diagonal the models differ in their classifications. Especially the differences among the models in their classifications of underperforming schools are interesting for accountability purposes.

For the five types of value added models, a variable was constructed indicating whether the value added was significantly below zero, zero, or above zero. After that, the consensus or agreement between the five models was estimated by calculating the Kappa. Table 2.9 shows to what extent the different value added models agree in their assignment of schools to one of the three groups. For prevocational education (VMBO tl) the agreement between the net school effects varies between moderate agreement (κ =.444; p<0.001; N=67; type AA and type X) and substantial agreement (κ =.732; p<0.001; N=67; type A and type B) (Landis & Koch, 1977). Between most types of school effects the agreement is substantial. The agreement between the gross school effect and the other types of school effects is less strong, varying between slight agreement (κ =.104; p<0.052; N=67; gross school effect and type X) and fair agreement (κ =.395; p<0.001; N=67; gross school effect and type AA).

Table 2.8

Example of HAVO school classifications for the 'type A' and 'type B' value added models

		Type A			
		Underperforming schools	Average schools	Overperforming schools	Total
Type B	Underperforming schools	2 (4.1%)			2 (4.1%)
	Average schools	3 (6.1%)	39 (79.6%)	1 (2.0%)	43 (87.8%)
	Overperforming schools			4 (8.2%)	4 (8.2%)
	Total	5 (10.2%)	39 (79.6%)	5 (10.2%)	49 (100%)

For higher general secondary education (HAVO) the agreement between the types of net school effects is slightly less strong and varies between fair and substantial. The lowest level of agreement can be found between the type AA and type X school effects (κ =.340; p<0.001; N=49). The agreement between types B and X is the strongest for higher general secondary education (κ =.783; p<0.001; N=49).

For higher general secondary education a similar pattern is found for the agreement between the different types of net school effects and the gross school effect.

Differences in the classification of schools into the three groups can be partly explained by the models on which these school effects are based. At the school level more variance is explained by the models with more covariates at both the student and the school level. This is why these models show fewer schools in the less and more effective than average groups. Especially the student level covariates in the type A school effect model accounted for a substantial part of the variance on the school level. Also shifts of schools have been found, for example schools which were depicted by the simple models as 'not significantly different from average' and by the more complex models as 'significantly different from average'.

Table 2.9

Agreement (Kappa) between types of school effects for the school types 'pre vocational education' (VMBO tl) and 'higher general secondary education' (HAVO)

	VMBO tl			
	Type 0	**Type AA**	**Type A**	**Type B**
Type AA	.395***			
Type A	.272***	.631***		
Type B	.210***	.676***	.732***	
Type X	.104	.444***	.490***	.632***
	HAVO			
	Type 0	**Type AA**	**Type A**	**Type B**
Type AA	.523***			
Type A	.400***	.716***		
Type B	.216**	.398***	.718***	
Type X	.199**	.340***	.530***	.783***

*p<0.05; **p<0.01; ***p<0.001 (two-tailed); VMBO n=67; HAVO n=49

2.4 Conclusion and discussion

The objective of the current study has been to assess the different operationalizations of value added models from both a conceptual and an empirical perspective to investigate which one(s) would fit the best in the accountability systems as used by authorities. We distinguished between five types of value added models, ranked by increasing complexity. All models adopt a multilevel approach to estimating school effects, but differ in their types of control variables. The first model is a basic model

without control variables (type 0), the second one is a model which controls for prior achievement (type AA), the third is one which controls for prior achievement and other student level covariates (type A), the fourth model controls for student level covariates and compositional covariates (type B) and the fifth one controls for student level covariates, compositional covariates, and other non-malleable school level covariates (type X). The consequence of the use of covariates plays an important role in the conceptual approach to the comparison of value added models. In the first place selection effects may occur, for example when some subgroups of students attend more effective schools. Only part of the school effect might be explained if the enrolment of some groups of students into particular schools is related to the potential benefit of this enrolment, for example the effectiveness of the school (Raudenbush, 2005). More research should be conducted to investigate the extent of these selection effects and how they can be positioned in the models for estimating school effects. Secondly, for some types of school effects covariates are suggested which are more closely related to the educational practices within the schools, for example compositional variables (Willms, 1992; Opdenakker et al., 2001; Cervini, 2005; Opdenakker et al., 2007). Controlling for these variables might also explain some of the other school effects by clarifying the joint effects of school composition and school policy and practices. These latter effects play a larger role in the suggested type B and type X school effect models, which include covariates at the school level. It is questionable whether the additional control variables in the type X school effect model have an added value in making the comparison of the outputs of schools more reliable, because value added indicators in educational accountability systems should identify possible weak schools, regardless of their denomination, structure, or location within the country.

In the empirical analysis the gross school effects as well as the effect types AA, A, B, and X were modelled. However, although the association and agreement in the classification of schools based on the different types of school effects might have provided some insights into the similarity of the types of school effects estimated, they cannot answer the question which type of school effect is the most appropriate to be used in an accountability system. A comparison between type AA and the other types of school effects proposed shows that the additional student and school level background variables have an important additional impact on top of the effect of prior achievement. Similar to the covariates in the type X school effect model it is questionable whether some of the covariates on the student level should be used when defining indicators for accountability systems. Using ethnicity or gender as control

variable implies that we expect and allow some subgroups of students to perform less successfully on their final examination, regardless of their prior achievement.

With regard to the average socio-economic status of students within the prevocational education segment the type B and type X models have shown significant compositional effects on the school level. Such compositional school effects should be handled with caution, as they might be an artefact of not having measured all non-cognitive characteristics on the student level associated with the students' relative progress (Nash, 2003).

Both for the prevocational and the higher general secondary education segments strong correlations have been found between the estimates of the different types of school effects. These results suggest that the different value added models represent a fairly similar construct. The agreement in the school classifications between the different operationalizations of the school effects is, however, less strong between the models. The explained variance at the school level can only account for part of the shifts of the groups. This measure of agreement shows that the choice of model has a large impact on the individual schools if this model would be used in an accountability system. Being incorrectly labelled as a less effective than average learning institution might have important implications for an individual school, such as, for example, intensified inspections. The modest level of agreement between the value added models and the implications for the individual schools suggest that in the development of these models for accountability purposes considerable thought should be given to the choice of covariates. We advise the value added models to include prior achievement, some indicators of the students' socio economic background, and compositional characteristics of student population.

Finally, a number of limitations of this study need to be considered. The analyses in the empirical part of our research are based on the VOCL'99 cohort. Although this is a representative sample of the schools in the Dutch secondary education system, differences in our sampling procedure compared to the national data might have affected the generalizability of our study, which only included the students who where initially selected for the sample. The effects of students outside the sample who changed schools to one of the sample schools at a later stage were not included in the analyses. In this sample all students started at the same time but graduated in different years. However, the estimated school effects on the Dutch student data included all students who graduate at the same time. Using the final examinations in successive years can cause problems in the estimation of the value added. Problems in the standardization of the scores on the final exam over successive years as well as

possible anomalies regarding equal performance and equal grades over the years, can cause bias in the value added estimates for the VOCL'99 data.

Secondly, the school where the students made their final examination was used in the analysis. Any effects of previous schools, in case of student mobility, were not specifically modelled. The might lead to underestimations of the school level variance in the present study.

Furthermore, this study have only described differences in conceptual meanings, operationalizations and empirical results of various .value added models No hard conclusions could be drawn about which value added model provides the most valid estimate of school effects on the basis of these results. Inspectorates of education may use the information obtained in this study in the development of indicators of school effectiveness. For educational accountability the transparency of the indicator is especially important. The better schools understand the value added indicator, the higher the chance that the indicator will be accepted as a measure of effectiveness. On behalf of the transparency of the indicator for stakeholders we suggest value added to be estimated by regression models instead of the more complex modelling of regression discontinuity approaches.

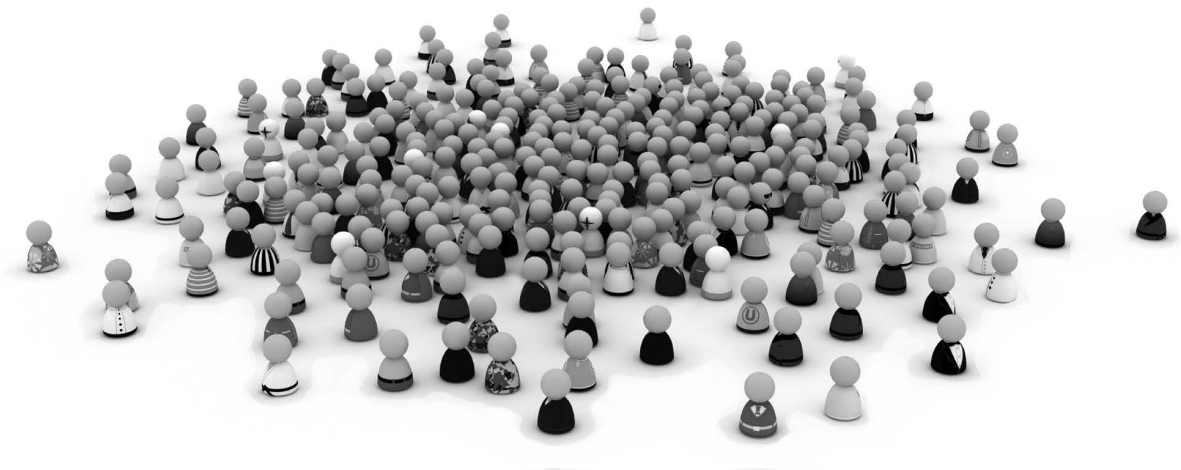
However, it should be kept in mind, however, that schooling is a complex process which cannot be captured in one measure of school effectiveness. Evidence from research into the consistency of school effects between subjects (Thomas et al., 1997b; Luyten, 1998; Ma, 2001) and differential school effects (Sammons et al., 1993; Thomas et al., 1997a; Veenstra, 1999; Gray et al., 2004) shows that a general value added indicator of school effectiveness masks all kinds of processes within the school that affect the progress of students. In an accountability system as in the Netherlands, which is based on identifying insufficient practices and schools as quickly as possible, a more sensitive system would be required in which multiple indicators of school effectiveness are simultaneously assessed, some of which are value added. In an accountability system the set of indicators to assess the quality of the school should include multiple indicators in a balanced system that refer to both cognitive and non-cognitive outcomes (Gray, 2004b), indicator referring to characteristics of effective schools or teachers and indicators that refer to effectiveness and efficiency. Combining multiple indicators of school quality decreases the possibilities of school to game the system (Figlio et al., 2002; Jacob, 2005; Cullen et al., 2006).

The focus of this paper has been on the function of identifying underperforming schools by education accountability systems and the usability of value added in such a system. Most accountability systems also share responsibility in school improvement,

besides the function of identifying underperformance. Using value added indictors for school improvement leads to additional requirements of these indicators. In order to provide feedback for improvement, at least more detailed information on value added for specific subgroups of students, specific subjects or departments should be given to schools.

Chapter 3
In Search of Value Added in the Case of Complex School Effects

This chapter is based on:
Timmermans, A. C., Snijders, T. A. B., & Bosker, R. J. (2012, In press). *In search of value added in case of complex school effects.* Journal of Educational and Psychological Measurement, DOI: 10.1177/0013164412460392



Abstract

In traditional studies on value added indicators of educational effectiveness students are usually treated as belonging to those schools where they made their final examination. However, in practice, students sometimes attend multiple schools and therefore it is questionable whether this assumption of belonging to the last school they attended can be made. Furthermore, the schools earlier attended by students might have long-term effects on their subsequent performance. Using data from Dutch primary and secondary schools, multiple membership models and cross classification multilevel models were estimated to explore the relationship between secondary schools, primary schools, and educational achievement simultaneously. Furthermore, the effects of student mobility and long-term primary school effects on the estimated value added of secondary schools has been explored. However, the long-term effects of primary schools did not change the estimated value added of secondary schools. On the other hand, allowing students to be a member of multiple secondary schools changed the estimated effectiveness of these schools especially for schools in the middle range of effectiveness.

3.1 Introduction

Value added indicators of school effectiveness are increasingly used in educational accountability systems to compare schools in their effects on students' achievement. Value added indicators promise to distinguish the schools' effectiveness more fairly by making a proper correction for differences in intake of students. Value added can be defined as "an indication of the extent to which any given school has fostered the progress of all students in a range of subjects during a particular time period in comparison to the effects of other schools in the same sample" (p.24) (Sammons et al., 1997). Usually the students' prior educational achievement, student background characteristics, and compositional variables of the student population are used as covariates in statistical models to account for differences in intake of students between secondary schools in order to achieve a fair comparison. The statistical techniques to estimate these value added indicators have been developed since the 1980's (Raudenbush et al., 1986; Aitkin et al., 1986; Willms et al., 1989; Raudenbush et al., 1995; Hill et al., 1996; Goldstein, 1997; Hill & Rowe, 1998). The assumption behind these value added indicators is that the effectiveness of schools can be considered as a latent trait, which can be measured through the performance of students within schools, just as estimating a latent trait in students can be achieved through a careful analysis of items.

Making a fair and valid comparison between the effectiveness of schools could also mean the correction for effects of previously attended schools, besides the usual covariates in the analysis of value added. Student mobility between secondary schools or long-term effects of primary schools are examples of previously attended education that might influence a students' performance during secondary education. Student mobility across schools and long term effects of primary schools traditionally are not incorporated in the estimation of value added of secondary schools. In the traditional value added models the students are treated as belonging to just one secondary school; the school where they did their final examination. Therefore, estimated value added derived from these traditional models might include influences of previously attended schools. Ignoring effects of previous education in the estimation of value added might lead to bias in the estimation of the latent school effectiveness trait. Further developments of methods and software made it possible to refine the value added indicators of school effectiveness and to allow for modelling educational data with

imperfect hierarchical structures (Hill & Goldstein, 1998; Browne, Goldstein, & Rasbash, 2001; Fielding & Goldstein, 2006).

The association of student mobility (Engec, 2006; Strand & Demie, 2007) as well as long-term effects of primary schools (Goldstein et al., 1997; Pustjens, Van de Gaer, Van Damme, Onghena, & Van Landeghem, 2007) with student achievement in secondary education has been studied before. However, research on the effects of these phenomena on the estimation of school effects in secondary education is rare (Leckie, 2009). The present study investigates the degree to which model specifications with respect to student mobility and long term primary school effects influence estimates of secondary school effects by means of multiple classification and multiple membership multilevel modelling (Browne et al., 2001; Fielding et al., 2006). The validity of the estimated value added of secondary schools is our main interest of this study rather than the phenomena of student mobility and long term primary school effects themselves. Modelling these imperfect hierarchical structures might have important consequences for the development of value added indicators for purposes of educational accountability. The following research question will be addressed in the remainder of this study. To what extent does modelling student mobility and long term primary school effects lead to differences in estimated value added indices in secondary education?

The following sections will give a brief overview of literature concerning effects of student mobility and long term primary school effects on student achievement during secondary education and the relation with the estimation of value added for secondary schools. In this section, a brief overview will be given of the methodology and results from previous studies. After that, the data, variables and analytical strategy are described including a short description of Dutch secondary education. Data from schools in Dutch secondary education will be used as an example for estimating value added indicators while controlling for student mobility and long term primary school effects. In the final section, the empirical results will be presented and discussed.

3.1.1 Multiple membership models and student mobility

Student mobility, also known as student transfer or school mobility, is defined as the movement of students between schools or educational institutions, once or multiple times, at other times than the normal age at which the students start or finish their education at a school (Strand et al., 2007). In this definition, the change from a primary school to a secondary school is not a part of student mobility. Student

mobility is known for its negative effects on student performance both on reading and mathematics (Temple & Reynolds, 1999; Mehana & Reynolds, 2004; Engec, 2006; Strand et al., 2007; Hattie, 2009).

The mobility of students between schools is mentioned as one of the many problems in estimating value added (Goldstein, 2001; Roeleveld, 2003b; Keeves et al., 2005). In traditional value added indicators of school effectiveness the students are treated as belonging to the school where they do their final examination. In these traditional value added models, all effects of previously attended schools are attributed to the final school regardless of whether the student changed schools. The regression formula for the hierarchical model for estimating traditional value added indicators in which student *i* (level 1) is nested within the final secondary school *j* (level 2) is given below in formula 1. In this formula, Y_{ij} is the dependent variable, usually test or examination scores at the end of a formal stage of education are used in the estimation of value added. The average performance of students in the sample is represented by the intercept, γ_0. Usually a set of control variables is included to explain parts of the variability in Y_{ij} at the student and school level. Control variables are depicted by $\gamma_h x_{hij}$ in the formula. In such models the residuals at the school level (U_{0j}) and the student level (R_{ij}) are assumed to be independent and with a population mean of 0 and a constant variance. The assumption behind this traditional value added model is that after a careful correction of differences between schools in their intake of students, the remaining between-school variance reflects differences between schools in effectiveness. The residual at the level of the secondary school is then considered to be the estimate of a schools' value added.

$$Y_{ij} = \gamma_0 + \sum_{h=1}^{q} \gamma_h x_{hij} + U_{0j} + R_{ij} \qquad (1)$$

First attempts were undertaken by Goldstein (2007) and Leckie (2009) to model student mobility in the estimation of value added indicators through the use of multiple membership multilevel models. In multiple membership models students (level 1) can be nested in multiple schools for secondary education (level 2) (Browne et al., 2001). The effects of the multiple secondary schools on the students' progress can be weighted (w_{ih}), for example, by the time the student attended the school. In formula 2, the multiple membership model is presented. In the notation chosen here, unlike in usual multilevel models, the index *i*, denoting the student, is not regarded as being nested in some higher-level unit, so that the values of *i* may range from 1 to the

total number of students in the data set. In a traditional multilevel model with students strictly nested in schools, for each student there is only one school h with $w_{ih} = 1$, and for all others $w_{ih} = 0$. In this case, the secondary school last attended will get the full weight, while all other schools get zero weight. In multiple membership models this is not the case, but in practice still most values of w_{ih} will be 0.

$$Y_i = \gamma_0 + \sum_{h=1}^{q} \gamma_h x_{hi} + \sum_{h=1}^{N} w_{ih} U_{0h} + R_i \qquad (2)$$

In a study in British primary education, Key Stage 2 (comparable to 3rd until 6th grade in the United States), a multiple membership cross-classified model was used for the combined analysis of the effects of student mobility and primary school attended (Goldstein et al., 2007). In a comparison a correlation was found of .98 between a model that took account of student mobility and prior education model and the traditional value added model. Goldstein et al. therefore suggest that ignoring student mobility and effects of earlier education does not appear to alter the rankings of schools on their posterior value-added estimates. However, they also show that ignoring student mobility leads to a downward bias of the estimated variance at the school level.

In a study in British secondary education, between Key Stage 2 and General Certificate of Secondary Education (GCSE, comparable to 7th until 11th grade in the United States) the effects of both student mobility and long term primary school effects on the estimation of value added of secondary schools are investigated (Leckie, 2009). In this study, primary schools were included as a crossed random effect. Similar to the findings of Goldstein et al. (2007), Leckie showed that the schools appear to be more important if student mobility is modelled through multiple membership models. In other words, ignoring student mobility leads to an underestimation of the between-secondary-school variance and the between-neighbourhood variance.

3.1.2 Multiple classification models and long term effects of previous education

Long term primary school effects on the learning progress of students during secondary education can be analysed through the use of multiple classification models. In multiple classification models students can be nested in groups on more than one dimension (Fielding et al., 2006). In these kinds of models students are both nested

within secondary schools and within a primary school. The random factors secondary schools and primary schools can both be seen as populations of interest. Primary and secondary schools are called crossed random factors, because not all students from the same primary school attend the same secondary and not all students from the same secondary school attended the same primary school.

In formula 3, a multiple classification model is presented for students *i*, nested within secondary schools *j*, and also nested within primary schools *k*. Secondary schools and primary schools are crossed random factors in this type of models. Compared to the traditional value added model in formula 1, the random effect of the primary school, indicated by W_{0k}, is just added to the formula. The usual assumption made is that the primary school effects are independent of the other random effects. The interpretation of the primary school effects is similar to other random effects, namely as representing the part of the variability in the dependent variable that is accounted for by primary schools. Similar to the previous models, covariates can be included in the model to control for differences in intake of students between secondary schools. After the inclusion of covariates indicating the performance or ability of students at the end of primary education, the residual between-primary school variance is assumed to reflect long term effects of primary schools.

$$Y_{i(j,k)} = \gamma_0 + \sum_{h=1}^{q} \gamma_h x_{hij} + U_{0j} + W_{0k} + R_{ij} \qquad (3)$$

The small body of literature concerning long term effects of previous education shows consistent small long term effects of previous schools on the subsequent performance of students (Tymms, 1995; Sammons, Nuttall, Cuttance, & Thomas, 1995b; Goldstein et al., 1997; Tymms, Merrell, & Henderson, 2000). However, results concerning the persistence of primary school effects during secondary education are inconsistent (Bressoux & Bianco, 2004; Creemers, Kyriakides, & Sammons, 2010). Differences between the results of the studies might arise from methodological differences between studies with respect to the inclusion of the teacher level or the department level and the period over which the long-term primary school effect is measured.

Small effects of primary schools, persisting during the entire duration of secondary education were found in British secondary education (Sammons et al., 1995b; Goldstein et al., 1997). Small long term effects of primary schools on student

achievement in secondary education were found in Flanders (Snijders et al., 1999; Pustjens et al., 2007). However, the small long term effects of primary schools on performance of students in secondary education in Flanders decreased rapidly during the first years of secondary education (Pustjens et al., 2007).

The inclusion of schools for primary education as a crossed random factor in a multiple classification model for the analysis of the effectiveness of secondary schools led to a great reduction in the estimated between secondary school variance in British secondary education (Goldstein et al., 1997). Students from effective primary schools also tend to do well at the end of secondary school. Serious distortions of the results of estimated value added might appear when no adjustments are made for previous education in the estimation of value added in secondary education. Because of these long term primary school effects, it was suggested that adjustments should be made not only for prior achievement but also for all previous education to find a better estimation of both short- and long-term school effects (Kyriakides et al., 2008).

3.2 Methods

3.2.1 Subjects

The data used here were collected as part of a national longitudinal study in secondary education in the Netherlands, the "Cohort Studies in Secondary Education" (Dutch abbreviation: VOCL). The study concerned students who entered the first grade of Dutch secondary education in the Netherlands (comparable to the 7th grade in the United States) in the year 1999, also called the VOCL'99 cohort. The total cohort consists of a sample of approximately 20,000 students. This sample has been considered as representative of the schools and students in the Dutch secondary education (Kuyper et al., 2003b). The data in the VOCL'99 cohort were derived from several sources and on several occasions (Kuyper et al., 2003a).

For the current study we selected a subsample from the VOCL'99 cohort based on the following criteria: identification variables had to be available at the student level, secondary school level, the primary school attended and the students made their final national examinations in the pre-vocational secondary education theoretical track (VMBO tl). Furthermore, as a result of data requirements for a correct estimation of the multiple membership models only those mobile students were included for whom examination results were available for the delivering and receiving secondary school. Both students who finished secondary education in the nominal time (4 years) and

students who lagged behind one year (5 years) were included in the sample. The Dutch secondary education system consists of multiple differentiated school tracks, for which the students are selected at age twelve on the basis of their scholastic aptitude. The VMBO theoretical programme is one of the four year vocational programmes preparing students for senior secondary vocational education. The subsample consists of 3658 students in 185 secondary schools. Due to student mobility the number of secondary schools is more than the original sample of 100 secondary schools. These students stem from 1292 different primary schools. The number of feeder primary schools for one secondary school ranged from 1 to 75 with a mean of 9 schools.

In this subsample only student mobility within the VMBO tl track was allowed. Within the VMBO tl track 94% of the students (3438) did their final national examination in the same school where they started their school career in secondary education. The remaining 6% of students was mobile during secondary education at least once. Of this group 213 students attended two secondary schools and 5 students attended three secondary schools. These five students who attended three schools all changed schools after the first year in secondary education.

3.2.2 Variables

Outcome variable The overall mean score on the final national examination was used as the dependent variable in the multilevel regression analysis, which ranges between 3 and 9. For the majority of students, who finished secondary education in the nominal time, results from the final nation examination in spring 2003 was used. For the students who lagged behind one year the scores on the final national examinations of spring 2004 was used.

Explanatory variables Halfway the seventh grade the "Cito-entry" test took place. The cito-entree test was developed by CITO, the Netherlands Institute for Educational Measurement. This test contains the parts Dutch language, mathematics, and information processing. For our study, the total score on the test was used as an overall measure of prior achievement. The total test had a reliability (Cronbach's α) of .90 (Kuyper et al., 2003b) and the range of scores on this test was between 13 and 60 points.

Another indication of students' prior scholastic aptitude is given by the advice that primary school teachers provide to parents at the end of primary education. The

advice is stated in terms of the most appropriate school track for the student in secondary education. The advice consists of nine categories and can range between a more individualistic track in pre-vocational secondary education and the pre-university school track. In the analysis the advice of the primary school teachers is used as a continuous variable.

Socio-economic status was measured by the highest educational level completed by one or both of the student's parents. This variable consisted of six categories (coded as 2 - 7), ranging from only primary to post-graduate education. In the analysis socio-economic status was used as a continuous variable. Descriptive statistics of the covariates are presented in table 3.1.

Information about the ethnic origin of students was gathered by asking the parents in which country they were born. Students' ethnicity was operationalized as a dichotomous variable with the categories native (coded as 0) and minority (coded as 1). Only if both parents and the student were born in the Netherlands was the student considered to be indigenous; in all other cases, the student was considered to be a minority student.

Table 3.1
Descriptive statistics of variables used for the estimation of value added models

	Average	Standard deviation	N	%
Dependent variable				
Final achievement	6.40	0.67	3658	
Explanatory variables				
Prior achievement	35.45	7.37	3469	
Advice from primary school teacher	4.03	0.99	3262	
Socio-economic status	5.41	1.22	3459	
Ethnicity (native)				84.5
Ethnicity (minority)				15.5

3.2.3 Methods of analysis

Multiple imputation through multilevel chained equations

The VOCL'99 cohort contained many predictors of the students' final achievement. For almost all predictor variables scores were missing for some pupils. The often used method of listwise deletion of cases with missing values is wasteful of information

and can lead to biases in results (Graham, 2009) and therefore we employed the multilevel Chained Equations technique (Van Buuren, 2011; Snijders & Bosker, 2012) using all available information. This method is very flexible and results from simulation studies are promising (Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006).

In total there were 2,088 complete cases and 1,570 pupils with one or more missing values on the predictor variables. The missingness is mostly not strongly associated between variables. First an initial random imputation was done to obtain a first complete dataset, based on a suboptimal but reasonable imputation in which the multilevel structure was ignored. The continuous variables, prior achievement, age, intelligence, advice and socio-economic status were randomly imputed based on a multivariate normal distribution jointly with the completely observed measure of final achievement. The variables second language, gender, age, intelligence, and living in a problem neighbourhood functioned as auxiliary variables for the imputation. For the dependent variables and the predictor variables with missing values, the main relations were investigated using the provisionally imputed dataset using multilevel analysis. This led to the imputation models, using the following rules; 1) significant variables and group means of significant variables were included, 2) if *X*-mean was a significant predictor for *Y*, then *Y*-mean was included as predictor of *X*, 3) implausible predictors were dropped and 4) unimportant predictors were dropped for binary dependent variables to improve convergence. We constructed 25 datasets with imputed values. Results reported in the following tables are the syntheses of 25 analyses run on these imputed data sets. For parameter estimates and standard errors, the combination rules of Rubin (1987) were used. The imputation uncertainty between imputated datasets appeared small because the estimated coefficients of the control variables hardly differed from each other, when the 25 datasets are considered. The missing fractions range between .014 (ethnicity) and .102 (advice), which indicates that at most 10,2% of the information in a variable was lost because of the missingness. Deviances reported are averages across the 25 imputed data sets, and because of the Bayesian estimation method and the imputations the deviance differences are to be used cautiously as indications of relative model fit.

Multiple membership and multiple classification multilevel models

Besides the traditional value added analysis used in school effects studies, three alternative models will be analysed in which deviations from the strict hierarchical structure are allowed (Snijders et al., 1999). Prior achievement, socio-economic status,

advice and ethnicity are included as covariates in all value added models. In the second model the effects of student mobility are included in the analysis using a multiple membership (MM) model. In this multiple membership model the weights given to each school are based on the proportion time spent in each school. The weight is equal to one for all 3,438 students who didn't change schools during secondary education. For the remaining students the non-zero weights for individual schools vary between 0.2 and 0.8. The total weight for each student is one. An overview of the mobility and weights in the sample is presented in Table 3.2.

Table 3.2
Overview of mobility and weights for secondary schools

Weights per secondary school	Number of students	Percentages of students
1	3438	94.0 %
0.2, 0.8	31	0.85 %
0.25, 0.75	17	0.46 %
0.40, 0.60	15	0.41 %
0.50, 0.50	49	1.34 %
0.60, 0.40	26	0.71 %
0.75, 0.25	48	1.31 %
0.25, 0.50, 0.25	1	0.03 %
0.40, 0.40, 0.20	2	0.05 %
0.50, 0.25, 0.25	1	0.03 %
0.60, 0.20, 0.20	1	0.03 %

The third model simultaneously analyses the effects of primary schools on students during secondary education, through a multiple classification (MC) multilevel model (Hill et al., 1998). The final model takes effects of primary schools and student mobility into account when estimating value added analysis of school effectiveness, by means of a multiple classification multiple membership (MMMC) model (Fielding et al., 2006).

The estimation of multiple-classified and multiple membership models runs into important computational limitations in existing maximum likelihood approaches (Browne et al., 2001). All of the models in this study are therefore estimated using Markov Chain Monte Carlo (MCMC) based algorithms from the MLwiN 2.22 software package for multilevel modelling (Rasbash et al., 2009; Browne, 2009). Starting values for the fixed parameters are estimated from simpler models using a

maximum likelihood approach in MLwiN. In these models grand mean centering was applied for all continuous covariates.

3.3 Results

3.3.1 Modelling value added estimates of school effectiveness

In Table 3.3 the results of the empty models are presented for all types of value added models. From the traditional value added model, it can be seen that the average examination score is 6.36 and the total variance 0.467. Of this variance, 14% is associated with the secondary schools. Intraclass correlations of similar magnitudes were found in previous studies in Dutch secondary education (Luyten, 1998; Veenstra, 1999).

Table 3.3
Results from empty models

	Traditional model		**MM model***		**MC model****		**MMMC model*****	
	Par.	S.E.	Par.	S.E.	Par.	S.E.	Par.	S.E.
Fixed effects								
Constant	6.36	0.03	6.360	0.03	6.36	0.03	6.36	0.03
Random effects								
Crossed random effect:								
Primary schools					0.006	0.005	0.006	0.005
Level-two random effect:								
Secondary schools	0.067	0.014	0.064	0.014	0.066	0.014	0.065	0.014
Level-one variance:								
Students	0.400	0.010	0.401	0.009	0.395	0.010	0.395	0.010
Model fit								
Deviance		7025.7		7035.0		6977.4		6984.1

* MM: Multiple membership model for modelling student mobility
** MC: Multiple classification model for including primary schools as a crossed random factor
*** MMMC: Multiple Membership Multiple Classification model for modelling student mobility and primary schools simultaneously

In the MM model, in which students are allowed to be a member of multiple secondary schools, there is a marginal decrease in the between-school variance, going down from 0.067 to 0.064, and an increase in the deviance with 10 points. This indicates that the multiple membership model does not seem to get meaningfully closer to the data than the traditional multilevel model.

The results in which the available information on the primary schools previously attended by the pupils is taken into account are presented in the MC model. Of course the average examination grade remains the same, but now we see some small changes in the variance components. The variance between secondary schools marginally decreases to 0.066, and the within-school variance decreases somewhat as well, as now the primary schools take up a variance component of 0.006. The decrease in deviance is 7,025.7 – 6,977.4 = 48.3, highly significant in a chi-squared distribution with d.f. = 1. However, the covariates such as prior achievement (start secondary education or end of primary education) are not yet included in these models and therefore the variance on the secondary school level cannot be seen as representing net between school differences but rather represents the gross secondary school effects.

The results of the MMMC model in which student mobility and long term primary school effects are estimated simultaneously are not very different from those of the multiple classification model. The deviance of this empty MMMC model is slightly higher than for the MC model. This is possible because of the Markov Chain Monte Carlo algorithm, suggesting that the model may have converged incompletely, and that this model, being more complicated, is harder to estimate than the earlier estimated models.

For the results presented in Table 3.4 the predictor variables prior achievement, socio-economic status, advice and ethnicity were included in the models to estimate value added. The four predictor variables all have highly significant effects, indicating that pupils with higher entry test scores, with higher recommendations from their primary school teachers, and from more affluent families have higher average examination scores. Moreover, pupils from ethnic minorities have lower examination results than pupils from the Dutch majority group. The results of these fixed effects are consistent over the four value added models.

Most important, however, are the estimates of the variance components. Comparing the models with and without predictor variables, all variance components have decreased because of the inclusion of the predictor variables. For the multiple classification model, the between-pupils within schools variance decreases from 0.395 to 0.330. The between- secondary-school variance (0.034) is almost half its original

estimate (0.066), which also turns out to be the case for the between-primary-school variance: from 0.006 this decreases to 0.003. The remaining between-secondary-school variance indicates that secondary schools do appear to have a value added effect on pupil achievement measured at the final examination. But primary schools, given the achievement levels attained by pupils at the end of primary education and given their family background, have only a marginally lasting effect as measured four or five years later at the secondary school examinations. However, the decrease in deviance between the traditional value added model and the MC value added model 6350.0 – 6326.1 = 23.9, is still highly significant in a chi-squared distribution with d.f. = 1. The effects of the multiple membership modelling of student mobility on the coefficients of the model are even smaller after the inclusion of predictor variables.

Table 3.4

Results from the multiple multilevel models for estimating value added

	Traditional value added model		**MM value added model**		**MC value added model**		**MMMC value added model**	
	Par.	S.E.	Par.	S.E.	Par.	S.E.	Par.	S.E.
Fixed effects								
Constant	6.39	0.02	6.40	0.02	6.39	0.02	6.40	0.02
Prior achievement	0.032	0.002	0.032	0.002	0.031	0.002	0.031	0.002
Socio-economic status	0.068	0.010	0.068	0.010	0.068	0.011	0.068	0.011
Advice	0.054	0.010	0.053	0.010	0.054	0.010	0.053	0.010
Ethnicity	-0.071	0.028	-0.072	0.028	-0.071	0.029	-0.073	0.028
Random effects								
Crossed random effect:								
Primary schools					0.003	0.003	0.003	0.003
Level-two random effect:								
Secondary schools	0.034	0.008	0.033	0.007	0.033	0.008	0.033	0.007
Level-one variance:								
Students	0.333	0.008	0.334	0.008	0.330	0.008	0.331	0.008
Model fit								
Deviance		6350.0		6366.1		6326.2		6338.6

3.3.2 Comparing the traditional value added model with the multiple classification and multiple membership models

The specific aim of this study was to investigate whether modelling the effects of the various imperfect hierarchical structures would affect the estimated value added of secondary schools. Correlations among residuals, value added scores, of secondary schools for the various models are presented in Table 3.5. Despite the very small but significant long term effects of primary schools, the inclusion of the multiple classification in the model doesn't appear to change the estimated value added of secondary schools. When the ranks of secondary schools according to their value added are considered, the inclusion of primary schools as a crossed random factor leads to a maximum shift in ranks of ten places in rank order compared to a traditional value added model.

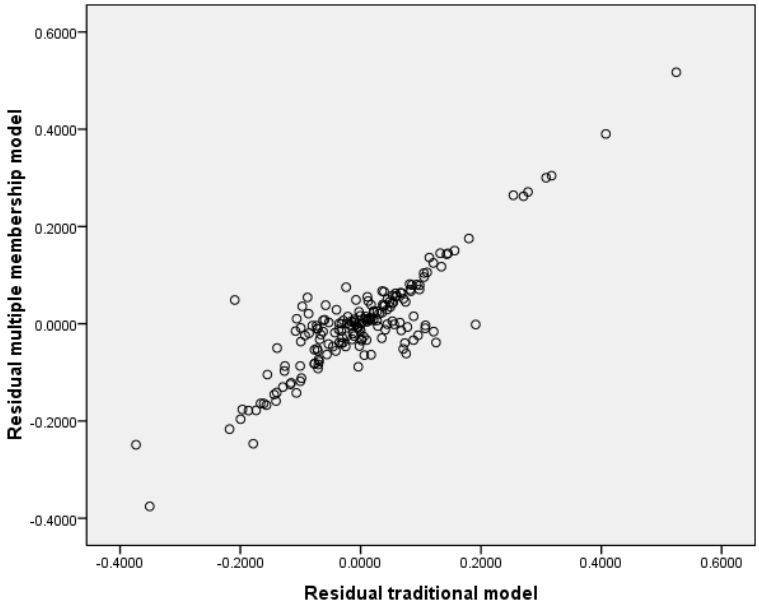
The results are somewhat different for the multiple membership model in which students are allowed to be a member of multiple secondary schools. A correlation of .88 was found between the estimated value added of secondary schools in a traditional model and in a multiple membership model. A scatterplot of the estimated value added in a traditional model and a multiple membership models is presented in Figure 3.1. From this scatter plot one can see that especially the schools in the middle range differ with respect to their value added for both models. The estimated value added of the most and least effective schools in the sample is relative stable over both models. If schools were compared in ranks for the traditional and multiple membership model over 50% of the schools changes ten places on the rank order or more.

Table 3.5

Correlations between school level residuals from the various multilevel models for estimating value added

	Traditional model	**MM model**	**MC model**
MM model	.880*		
MC model	1.00*	.879*	
MMMC model	.881*	1.00*	.880*

Figure 3.1
Scatterplot of value added estimates of secondary schools derived from a traditional model and from a multiple membership model.



3.4 Conclusion and discussion

The main focus of this study was to investigate the degree to which model specifications with respect to student mobility during secondary education and long term effects of primary schools influence the estimation of value added for secondary schools. Traditional studies in school effectiveness research and several educational accountability systems apply multilevel models in which the students are strictly nested within schools. However, there is some evidence of long term effects of previously attended schools (Pustjens et al., 2007) and students may attend more than a single school during a formal period of schooling. These long term effects and student mobility might bias the estimated value added of secondary schools if they are ignored in the analyses.

Value added indicators, which are frequently used in educational accountability systems, should be valid but not unnecessarily complicated, for the reason that the indicator should be as transparent as possible. Only if the complex modelling of

student mobility and effects of attended primary schools have important effects on the estimated value added of secondary schools, these models should be applied in educational accountability systems. Otherwise, if these complex models do not alter the estimated value added importantly more simple models are preferable.

In the current study we found very small but significant long term effects of primary schools on the performance of students on their final examination in secondary education. Based on previous literature these small effects are not surprising. However, it is important to keep in mind that the dependent variable in the model was student achievement at the end of secondary education measured four or five years after the students left the primary schools. Even though there appeared very small primary school effects, the inclusion of the multiple classification of primary schools with secondary schools didn't alter the estimated value added of secondary schools. These results are in contrast to the findings of Leckie (2009) that showed that including long term effects of primary schools in the analysis of value added of secondary did change the estimated secondary school effects.

Allowing students to be a member of multiple secondary schools however did appear to have an effect on the estimated value added of secondary schools. A strong, positive correlation was found between a traditional value added model and a multiple membership model. However, over 50% of the schools changes more than 10 places in the rank order. Differences between the estimated value added of the traditional model and the multiple membership model imply that student mobility should be included in the analysis. Somewhat stronger correlations were found between the results of a traditional value added models and a model in that took account of student mobility in British secondary education (Leckie, 2009). In this current study, the inclusion of multiple membership in the model changes the estimated value added especially for secondary schools in the middle range. Estimated value added for the most and least effective schools seemed rather stable over the models. Most educational accountability systems are designed to identify potential underperforming schools. The relative stable effects of value added estimates for the weakest schools over different models implies that traditional value added models seem sufficient in identifying underperformance.

A number of limitations of the data and the models applied in this study should be considered when interpreting the results. In the first place, only data from one of the tracks in a differentiated educational system is used in this study for the analyses of long term primary school effects and student mobility. This can be regarded as a relatively homogeneous population. The small differences between the various value

added models might partly be due to this relatively homogeneous character of the sample. Furthermore, results from one track cannot easily be generalized to other tracks, as tracks differ in length, content, level and possibilities for between track mobility. The effects of long term primary school effects and student mobility on the estimated value added of secondary schools might depend heavily on these track characteristics.

Secondly, only student mobility within the VMBO tl school track was estimated in this study due to data requirements for estimating multiple membership models. Examination results had to be available for both the delivering and receiving school. Especially in strongly differentiated educational systems, such as the Dutch secondary education, where not all schools provide education in all tracks there can be considerable student mobility between tracks. In such a differentiated educational system, the multiple membership models only partly resolves the student mobility problem on the estimation of secondary school effects because of the data requirements and the subsequent underestimation of student mobility in differentiated educational systems.

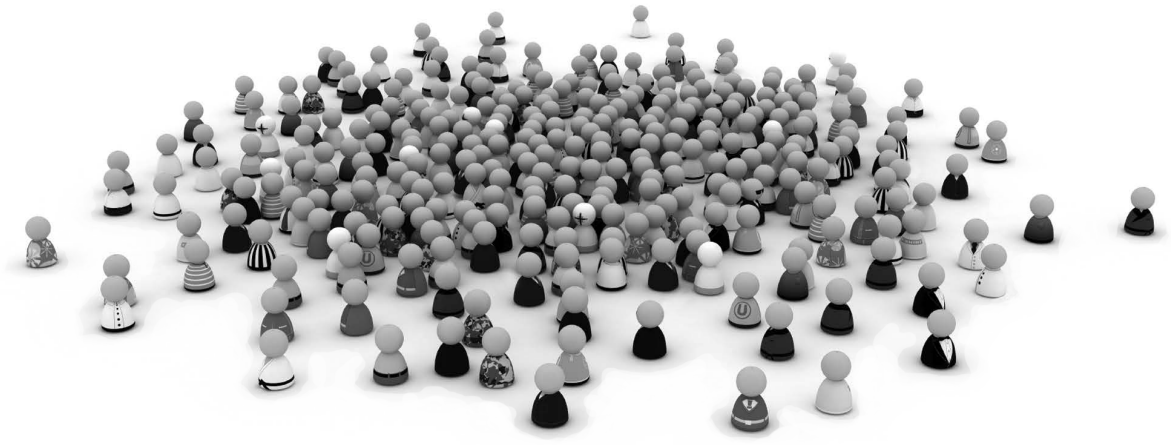
Furthermore, in multiple membership models lower level units are allowed to be a member of multiple units at the higher level, in this case a student can be a member of multiple secondary schools. The multiple membership models however cannot account for the order in which the students attended the secondary schools, which might cause bias in the estimated value added. The effects of a secondary school, in the case of student mobility, might be passed on to the subsequent school. Alternative weighting options in multiple membership models can be explored in order to assess the impact of ordering of schools on student performance, a combination of time spent in school and order of schools might be considered. In a previous study Britisch secondary education, the time spent in schools as weighting for secondary schools showed the best fit with the data (Leckie, 2009). Furthermore, the data used in this study didn't allow for a very precise determination of the weights for the multiple membership models. Which school the student attended was only registered once every year. Therefore, we might miss some schools if students attended them very briefly and we might misestimate the time spend in schools by students, because we have only one measurement per year. Despite these imprecise measurements of time spent in schools, the study nevertheless shed some light on the effects of student mobility on the estimation of value added indicators for secondary schools.

For future research we suggest to assess the effects of student mobility and long term effects of primary schools on the estimation of secondary school effectiveness

on larger datasets, for example national student databases. Furthermore, the effects of primary schools might also be investigated by including the average final achievement per primary school as a predictor in the analyses of value added for secondary schools, since this is an indicator of observed quality.

Chapter 4

Educational Accountability Based on the Cognitive and Non-cognitive Performance of Students: A Value Added Approach



Abstract

In this chapter, effects of schools on both cognitive and non-cognitive outcomes in Dutch secondary education, in the context of educational accountability, are explored by means of multivariate multilevel analysis. The sample for this study consisted of 10,849 students in 82 schools. Our study confirms that the relative influence of schools is much higher for the cognitive domain than for the non-cognitive domain. Moderately strong correlations were found between school effects on the perceived classroom climate and school effects on mathematics and language achievement. However, correlations of school effects on mathematics and language with achievement motivation are small and not significant. The evidence of a single underlying dimension of school effectiveness is limited. Furthermore, the implications of these results for the development of educational accountability systems are considered.

4.1 Introduction

One of the main criticisms on Educational Effectiveness Research (EER) is its narrow focus on disentangling effects of schools, classes or teachers on *cognitive* performance of students (Coe et al., 1998; Reynolds & Teddlie, 1999a; Teddlie et al., 2000a; Luyten, Visscher, & Witziers, 2005). The main reason for the focus on cognitive performance is that achievement in the cognitive domain has been considered as more important than achievement in the social or other domains, according to without any convincing reasons (Roede, 2001). Also, this focus on cognitive achievement might be exacerbated due to cognitive performance being relatively easily measured in tests. In many countries, large scale datasets on these test scores are easily available. However, not only the non-cognitive domain has been neglected in EER, but the same holds for some difficult to measure aspects of the cognitive domain, such as speaking foreign languages or presenting tasks.

Performance indicators used in educational accountability have this same focus on cognitive student performance and test scores. Examples are accountability systems in the Netherlands (Inspectorate of Education, 2009), England (Ray, 2006; Ofsted, 2010) and in all states of the US, for example Tennessee (Sanders et al., 1994; Sanders, 2003), Colorado (Betebenner, 2007; Betebenner, 2009) and Michigan (Lee & Weimer, 2002). The current set of indicators of school performance or quality in Dutch educational accountability, as developed by the Inspectorate of Education, has a strong focus on cognitive performance of students, which leads to many critiques from schools. These critiques are twofold, namely critiques on the accuracy and validity of indicators within the cognitive domain on one hand and on the other hand the restricted focus on the cognitive performance of schools, even though schools may pursue many other goals than cognitive achievement in mathematics and language, such as social skills and citizenship. This restricted focus on cognitive performance is widely criticized.

The main focus of this current study is the estimation of value added based on multiple outcomes of education, both in the cognitive and non-cognitive domain. This study investigates the uni- or multidimensionality of educational effectiveness, the magnitude of school differences and explores possibilities to use non-cognitive

outcomes for educational accountability. The following research questions will be answered in this study:

1. How large are the gross and net differences between secondary schools on cognitive and non-cognitive outcomes?
2. How strong is the association between the gross school effects and value added estimates of secondary schools for cognitive and non-cognitive outcomes?

Especially for the context of educational accountability it is important to estimate differences between secondary schools in terms of value added, which is considered the most appropriate indicator to compare school in their performance (Meyer, 1997; Bosker et al., 2001; Schagen et al., 2003; OECD, 2008). We also chose to use data from schools in the Netherlands because of the interesting variation in type of schools in the Netherlands. An overview of the Dutch educational system in presented in Appendix 1.

The structure of the paper is rather basic. The following section will give a brief description of the discussion on outcomes in education and an overview is given on previous results of estimation of school effects within the non-cognitive domain. The second section gives an overview of the data and methods used in this study. In the third section, the empirical results will be presented. Implications of the results for both educational effectiveness research and educational accountability will be discussed in the last section of this paper.

4.1.1 Outcomes of education

In an ideal situation, the set of outcome measures in EER and educational accountability covers all skills and knowledge that are demanded by society (Meyer, 1997) and measures all key qualifications of educational achievement of schools (Hill et al., 1996; Coe et al., 1998; Peschar et al., 2001). This means that outcome measures for school accountability should include measures on cognitive skills, but also on the development of personal, affective en social skills (Peschar et al., 2001). Moreover, several studies have shown that non-cognitive outcomes of education are crucial for the achievement of students within schools, as well as personal functioning and participation in society (Mortimore, Sammons, Stoll, Lewis, & Ecob, 1988; Solomon, Watson, Delucchi, Schaps, & Battistich, 1988; Freiberg, 1996; Lewis, Schaps, & Watson, 1996). Multidimensional measures of school effectiveness, incorporating both cognitive and non-cognitive outcomes are therefore advocated (Teddlie et al., 2000a),

because non-cognitive outcomes are important and seen as educational aims in themselves. In addition, Roede (2001) suggests school and classroom climate, social development of students as possible examples of other outcomes of education for further research.

Six dimensions of key qualifications have been formulated for cognitive and non-cognitive outcomes of education for an optimal adaptation of students to society, future education and jobs (Van Zolingen, 1995; Van Zolingen & KLaassen, 2003). An overview of the key qualifications is given in Table 4.1. Both educational effectiveness research and educational accountability have focussed mainly on the first dimension of key qualifications.

Table 4.1.
Overview of key qualifications

Key qualification	Description
General-instrumental	Knowledge and skills with a permanent character and that can be applied in many situations and interdisciplinary knowledge. (For example basic skills in mathematics, language, reading, ability to plan work, ability to handle information)
Cognitive	Thinking and acting. (For example, identifying and solving problems, abstract thinking, learning to learn, intellectual flexibility)
Personality	Individual behaviour. (Such as sense of responsibility, accuracy, confidence, creativity, willingness to achieve, coping with stress)
Socio-communicative	Communicating skills (For example expressing oneself orally and in writing) and the ability to work together others (For example, social skills, solidarity and empathy)
Socio-normative	Ability to adapt to the corporate culture (Such as, identification, dedication, knowledge of an organization, complying with safety measures)
Strategic	Emancipatory behaviour. (For example, taking an active part in decision making, dealing critically with choice and effects they have)

The current societal and political opinion in the Netherlands is that schools should be autonomous and have their own responsibilities in providing and organizing their education (Ehren et al., 2005). This has been determined in Article 23 of the Dutch Constitution. This freedom implies that everyone can start a school with public funding, as long as there is a minimum number of students and the school fulfills minimal requirements. However, minimum standards and exam requirements are formulated that every school should meet. Due to this relative freedom, the Dutch educational system shows a wide variety of schools based on their religious beliefs and educational concepts. A number of schools articulate the importance of a wider development of students more clearly, such as Montessori or Waldorf schools. Waldorf schools are an example of schools that wide focus on the development of students. In these schools the cognitive development of students is equally as important as their creative, social and emotional development (Steenbergen, 2009). However, all schools agree with the importance of the development of other than cognitive skills for their students.

However, in developing indicators for the quality aspects of education the inspectorate of education moves on a small line between the freedom that schools have to develop their education and to determine their own focus and the necessity of boundaries on this freedom for accountability by governments, as stated in article 23 of the Dutch Constitution (Onderwijsraad, 1999; Onderwijsraad, 2002; de Nationale ombudsman, 2009). Each indicator developed by the Inspectorate of Education for the use in educational accountability reduces the freedom of the school. Therefore, additional quality indicators for the non-cognitive outcomes of education might lead to a more complete picture of a schools' effectiveness in exchange of a reduction of freedom and autonomy.

4.1.2 School effects in the non-cognitive domain

A number of studies investigated the effects of schools or teachers on non-cognitive outcomes of education. However, it is difficult to compare these studies because the non-cognitive outcomes are defined in several ways and very differing sets of control variables were used for each study. In general, previous research has shown rather small effects of schools and classes on non-cognitive outcomes in primary education (Knuver, 1993; Hofman, Hofman, & Guldemond, 1999) and secondary education (Opdenakker & Van Damme, 2000; Van Landeghem, Van Damme, Opdenakker, De Fraine, & Onghena, 2002; Konu, Lintonen, & Autio, 2002; Van Damme, De Fraine, Van Landeghem, Opdenakker, & Onghena, 2002; Snijders et al., 2012). The effects of

schools and classes on non-cognitive outcomes are considerably smaller than school effects on the cognitive achievement of students (Opdenakker et al., 2000; Van Landeghem et al., 2002; Gray, 2004b; Van Damme et al., 2006). These findings question the possibilities of schools or classes to improve student outcomes in non-cognitive domains.

Moreover, there is some evidence that effective schools on the cognitive domain are not necessarily effective schools in non-cognitive domains (Thomas, Smees, MacBeath, Robertson, & Boyd, 2000; Gray, 2004b). This implies that evidence for a single underlying dimension of school effectiveness is limited. Especially in primary education, school effects in the cognitive and non-cognitive domain are weakly positively related and may be independent (Knuver, 1993; Teddlie et al., 2000a). In secondary education, the association between school effects on the cognitive and specific aspects of the non-cognitive domain appears small but significant (Thomas et al., 2000).

Furthermore, control variables that explain differences in cognitive achievement between students and schools have limited explanatory power for the non-cognitive domain. Prior cognitive achievement is usually the best predictor of the students' achievement at the end of an educational stage. Yet, specific control variables for the non-cognitive domain need to be identified. In the study of Thomas and colleagues (2000), the association between prior and final attitudes test with a two year gap ranged between .22 and .50. Although these correlations are smaller than correlations usually found for the cognitive domain, the prior attitudes appeared to be the best predictors of students' final attitudes.

4.2 Method

4.2.1 Dataset

The data used in this empirical study were collected as part of a national longitudinal study in secondary education in the Netherlands, the "Cohort Studies in Secondary Education" (Dutch abbreviation: VOCL). The data concerns students who entered the first grade of Dutch secondary education in the Netherlands (comparable to the 7th grade in the United States) in the year 1999, also called the VOCL'99 cohort. The total cohort consists of an initial sample of approximately 20,000 students in 100 schools. This sample has been considered as representative of the schools and students in Dutch secondary education (Kuyper et al., 2003b). The data in the

VOCL'99 cohort were derived from several sources and on several occasions (Kuyper et al., 2003a).

For the current study we selected a subsample from the VOCL'99 cohort based on the following two criteria: 1) identification variables had to be available at the student level and school level, and 2) at least scores on one of the dependent variables, measured in the third year of secondary education, had to be available. This resulted in the use of a subsample of 10,849 students in 82 schools for secondary education. These students stem from all school tracks in the Dutch secondary educational system.

4.2.2 Cognitive and non-cognitive outcomes

In this study, we use two cognitive and two non-cognitive outcome measures based on the availability of possible outcomes of education in the VOCL'99 cohort. The cognitive outcome measures are achievement in Dutch Language and achievement in Mathematics. These indicators are available at the student level.

A first non-cognitive outcome measure is *classroom climate* as perceived by the student. Classroom climate reflects the social skills of students in a classroom, as measured by helping, trusting, and being nice towards each other, treating students and teachers fairly, accepting fellow students for who they are and whether the climate is the class is friendly, noisy or loud and whether there is a lot of calling names going on. This is an educational outcome that can be seen as a part of the *socio-communicative* dimension of the key qualifications, as described in Table 4.1. It reflects both the communicating skills and the ability of working together from the socio-communicative dimension. Besides the importance as an outcome variable, having a good classroom climate can be seen a condition for creating an environment with a focus on teaching and learning and is therefore associated with student achievement (Haertel, Walberg, & Haertel, 1981; Sammons et al., 1995a; Hattie, 2009; Teodorovic, 2011).

A second non-cognitive outcome measure we use, is *achievement motivation*. Achievement motivation refers to the tendency of a person to want to achieve (Atkinson & Reitman, 1958). This educational outcome of education is a part of the *personality* dimension of the key qualifications, as described in Table 4.1. Achievement motivation can be interpreted as 'motivation to learn'. It has not only a crucial role in students' learning during their school career, but also on the labour market and in their life-long learning. Achievement motivation is related to student achievement,

attainment and related to movements in students' educational career (Hattie, 2009; Hustinx, Kuyper, Van der Werf, & Dijkstra, 2009; Kuyper, Van der Werf, & Lubbers, 2010).

4.2.3 Design

The cognitive and non-cognitive outcomes are tested at the beginning of secondary education and in the third year of secondary education, which gives us prior and final scores on these measures and the possibility to estimate value added in the cognitive and non-cognitive domain. Especially for the context of educational accountability it is important to estimate differences between secondary schools in terms of value added, which is considered the most appropriate indicator to compare school in their performance (Meyer, 1997; Bosker et al., 2001; Schagen et al., 2003; OECD, 2008). For the outcomes in the cognitive domain we have used a mathematics and language test in the third year of education that was made by students in all school tracks and therefore we can estimate value added for a complete secondary school.

Table 4.2
Overview of variables used in this study

Variables	**Measurement occasion**		**Measurement level**
	7th grade	**9th grade**	
Language performance	Explanatory variable	Dependent variable	Student
Mathematics performance	Explanatory variable	Dependent variable	Student
Classroom climate	Explanatory variable	Dependent variable	Student
Achievement motivation	Explanatory variable	Dependent variable	Student
Socio-economic status	Explanatory variable		Student
Gender	Explanatory variable		Student
Ethnicity	Explanatory variable		Student
Second language	Explanatory variable		Student
School type		Explanatory variable	Student

4.2.4 Variables and instruments

Of focal interest in this study were the variables representing the performance of students in the third year of secondary education, this comparable to ninth grade in the United States. *Prior achievement* for the same outcomes, *socio-economic status*, *gender*, *ethnicity*, *second language* and *school type* functioned as covariates in the analyses where differences in effectiveness between schools are established. An overview of the

variables and the measurement occasion of these variables is presented in table 4.2. The variables and their instrumentation used in the analysis are discussed below.

Dependent variables

Achievement in Dutch language The Dutch language and reading test were originally developed by CITO, the Netherlands Institute for Educational Measurement. This test contains 34 multiple choice items with four answer categories for six text fragments. An internal reliability (Chronbach's α) of .75 was reported for the 34 items (Zijsling, Kuyper, Lubbers, & Van der Werf, 2005).

Achievement in Mathematics The Mathematics test originally developed by CITO, is administered in two versions of both 41 multiple choice items with four answer categories. Version A, with an internal reliability (Chronbach's α) of .78, was made by students from the higher school tracks (pre-university education, higher general secondary education and pre-vocational secondary education theoretical track). Students from the lower school tracks made the B version of the Mathematics test. An internal reliability (Chronbach's α) of .82 was reported for this version (Zijsling et al., 2005). Scores on both versions were made equivalent using OPLM.

Classroom climate The perceived classroom climate by students was measured using an 8 items scale with an internal reliability (Chronbach's α) of .83 and is largely based on previous developed tests (Veugelers & De Kat, 1998). An example of an item from the scale is: "In our class, students are nice towards each other". Students could state whether they agreed with this items using a five point likert-scale.

Achievement motivation Achievement motivation was measured using 9 items with four answer categories. An internal reliability (Chronbach's α) of .76 was reported for this scale (Zijsling et al., 2005). An example of an item from the scale is: "I would like to be the best student in my class".

Explanatory variables

Halfway the seventh grade the "cito-entree" test took place. The cito-entree test has been developed by CITO, the Netherlands Institute for Educational Measurement. This test assesses pupil achievement in various domains, such as Dutch language (Chronbach's α .74) and mathematics (Chronbach's α .83). The separate parts of the test were used as predictor variables in the analysis. Each of the separate parts of the test has a range between 1 and 20 points.

Table 4.3

Descriptive statistics of variables used for the estimation of value added models

	Average	Standard deviation	Percentage	N
Dependent variable				
Language scores	51.63	10.34		10,312
Mathematics scores	51.58	10.39		10,060
Classroom climate	3.54	0.64		9,573
Achievement motivation	2.57	0.51		9,507
Explanatory variables				
Prior achievement: Language	12.91	3.75		10,375
Prior achievement: Mathematics	12.78	4.40		10,371
Prior scores on classroom climate	3.64	0.62		10,340
Prior scores on achievement motivation	2.86	0.45		10,172
Socio-economic status	4.15	1.10		9,776
Ethnicity (minority)			14.9	10,786
Gender (girls)			52.1	10,849
Second language:				9,816
Dutch Language			77.6	
Bilingual			6.1	
Dialect			13.8	
Other language			2.5	
School type				10,814
Pre-university education			24.6	
Combined pre-university education and higher general secondary education			2.6	
Higher general secondary education			21.8	
Pre-vocation education theoretical track			28.6	
Pre-vocation education middle track			9.0	
Combined pre-vocation education middle and basic track			0.5	
Pre-vocation education basic track			10.7	
Pre-vocation education basic track with additional support			2.2	

For the non-cognitive outcomes a student questionnaire was administered in the seventh grade in which the two non-cognitive outcomes were included. Classroom climate was measured using an 8 items scale with an internal reliability (Chronbach's α) of .81 for the total cohort and is largely based on previous developed tests (Veugelers et al., 1998). This variable is used as predictor for the dependent variable classroom climate measured in the third year. Achievement motivation was measured using 16 items with four answer categories. An internal reliability (Chronbach's α) of

.80 was reported for this scale in the total cohort (Kuyper et al., 2003a). Achievement motivation measured in the first year is used as predictor for achievement motivation measured in the third year.

Socio-economic status was measured by the highest educational level completed by one or both of the student's parents. This variable consisted of six categories, ranging from only primary education to post-graduate. In the analysis, socio-economic status was used as a continuous variable. Dummy variables were created for second language, ethnicity, gender and school type. Descriptive statistics of the previous predictor variables are presented in table 4.3. For the second language, the only Dutch speaking group of students is used as reference group in the analysis. For school type, pre-vocation education basic track with additional support is used as reference group.

4.2.5 Methods of analysis

To analyse the cognitive and non-cognitive outcome variables simultaneously a multivariate multilevel model was estimated. Language achievement, mathematics achievement, classroom climate and achievement motivation are used as the dependent variables. These variables are standardized for the analysis. An elegant characteristic of this model is that students with missing scores on one or more dependent variables remain in the model (Snijders et al., 1999). Furthermore, the multivariate models provide the covariance between the estimated school effects for the cognitive and non-cognitive dependent variables. These covariances can be used to estimate the correlation between the estimated school effects.

This multivariate multilevel model is created in two steps. First, an empty multivariate model for the cognitive and non-cognitive dependent variables is estimated to investigate the variance accounted for at the school level, also known as a model for estimating gross school effects. Predictor variables were included in a second stage to estimate value added. Separate coefficients were estimated for the predictor variables on each of the dependent variables, resulting in different regression formulas for each of the dependent variables. In the models in which the predictor variables were included only students were included with complete records on the explanatory variables. The models are estimated using IGLS estimation methods from the MLwiN 2.24 software for multilevel modeling (Rasbash et al., 2009). In these models grand mean centring was applied for all continuous covariates.

4.2.6 Attrition

Due to missing values on one or more covariates 1,399 students of the original sample were lost from the analysis. This is 12.9 % of the original subsample. Socio-economic status (9.9%) and second language (9.5%) are the control variables with the largest amount of missing values. Comparing students who were included in the analysis and the students who were excluded from the analysis revealed some possible sources of attrition bias. Students included in the analysis come on average from more affluent families than students excluded from the analysis due to missing values (t=4.69; df= 880.22; p<0.001). A larger proportion from the students excluded from the analysis speaks a dialect or is bilingual compared to students included in the analysis (χ^2=12.95; df=3; p=0.005). Furthermore, a larger proportion of the group students excluded from the analysis are minority students (t=-9.54; df=2225.88; p<0.001). For gender we did not find signs of attrition bias.

Table 4.4

Results from gross school effect models for cognitive and non-cognitive outcomes

	Language		**Mathematics**		**Classroom climate**		**Achievement motivation**	
	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
Fixed effects								
Constant	-0.081	0.65	-0.029	0.66	-0.077	0.04	0.025	0.03
Random effects								
Secondary schools (τ^2)	0.333	0.05	0.343	0.06	0.107	0.02	0.037	0.01
Students (σ^2)	0.782	0.01	0.740	0.01	0.895	0.01	0.956	0.01
Intraclass correlation								
ICC	.299		.317		.105		.039	

-2loglikelihood: 224189.8; number of schools 82; number of students 10,849

* Significant at α=.05 (two tailed)

4.3 Results

Results from the empty multivariate multilevel model are presented in Table 4.4. Differences between schools obtained from this empty model can be interpreted as gross school effects. It is clearly visible from the intraclass correlations that the percentage of variance on the school level is much higher for the cognitive outcomes

than it is for the non-cognitive outcomes. For example, for mathematics almost 32% of the variance (ϱ = 37.01/(37.01+79.86)= 0.317)[3] is attributed to the level of the secondary schools, while for the achievement motivation only 4% of the variance (ϱ = 0.010/(0.010+0.249)= 0.039) is accounted for by the secondary schools. This means that schools are more homogeneous on non-cognitive outcomes than they are for the cognitive outcomes. Especially for the achievement motivation the differences between schools are very small. The amount of variance on the school level for the cognitive domain seems marginal larger in this study compared to results from a previous cohort in Dutch secondary education (Thomas, 2001). However, schools differ in their intake of students and therefore the unexplained variance from this model both on the student and school level include effects of different sources outside the practices and policies of the schools. These gross school effects cannot be considered to be most appropriate for education accountability for this same reason (Willms et al., 1989; Raudenbush et al., 1995; Timmermans, Doolaard, & De Wolf, 2011).

In Table 4.5, the results are presented in which several predictor variables are included to estimate the net school effects, or value added. For all outcome variables a positive significant relationship was found between prior and final achievement scores, also if other student characteristics were included in the model. Similar to the study of Thomas et al. (2000), the bivariate relationships between prior and final achievement scores are substantially stronger in the cognitive domain than in the non-cognitive domain, as shown in table 4.6.

[3] The intraclass correlation can be derived from the estimated variances with the following formula (Hox & Roberts, 2011): $\rho = \frac{\tau_0^2}{\tau_0^2 + \sigma^2}$

Table 4.5
Results of value added models for the cognitive and non-cognitive outcomes

	Language		Mathematics		Classroom climate		Achievement motivation	
	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
Fixed effects								
Constant	-0.936*	0.076	-0.783*	0.064	-0.403*	0.086	-0.090*	0.088
Prior achievement:								
Language	0.046*	0.003						
Mathematics			0.061*	0.002				
Classroom climate					0.369*	0.017		
Achievement motivation							0.774*	0.023
Student characteristics								
Socio-economic status	0.025*	0.008	0.025*	0.007	-0.006	0.010	0.006	0.011
Gender (girls)	0.152*	0.016	-0.078*	0.015	0.134*	0.020	0.093*	0.020
Ethnicity (non-Dutch students)	-0.104*	0.027	-0.102*	0.024	-0.009	0.033	0.103*	0.035
Second language:								
Biligual	-0.105*	0.036	-0.062	0.032	0.004	0.044	0.070	0.045
Dialect	-0.04	0.030	0.014	0.027	0.034	0.036	0.056	0.037
Other language	0.067	0.058	-0.023	0.051	-0.001	0.071	0.271*	0.074
School type								
Pre-university education	1.526*	0.074	1.584*	0.066	0.695*	0.085	0.103	0.089
Combined pre-university education and higher general secondary education	1.382*	0.091	1.476*	0.081	0.346*	0.113	0.188	0.116
Higher general secondary education	1.024*	0.071	1.028*	0.064	0.377*	0.084	0.125	0.089
Pre-vocation education theoretical track	0.587*	0.068	0.549*	0.061	0.137	0.083	-0.032	0.087
Pre-vocation education middle track	0.369*	0.070	0.376*	0.062	-0.001	0.086	-0.107	0.091
Combined pre-vocation education middle and basic track	-0.167	0.171	0.170	0.149	-0.597*	0.202	-0.292	0.192
Pre-vocation education basic track	0.102	0.069	0.163*	0.061	-0.171*	0.085	-0.157	0.090
Random effects								
Secondary schools (τ^2)	0.096	0.017	0.040	0.007	0.062	0.012	0.025	0.006
Students (σ^2)	0.543	0.008	0.434	0.007	0.768	0.012	0.815	0.013
Intraclass correlation								
ICC	.15		.08		.09		.03	

-2loglikelihood: 80,106.54; number of schools 82; number of students 9,450

* Significant at α=.05 (two tailed)

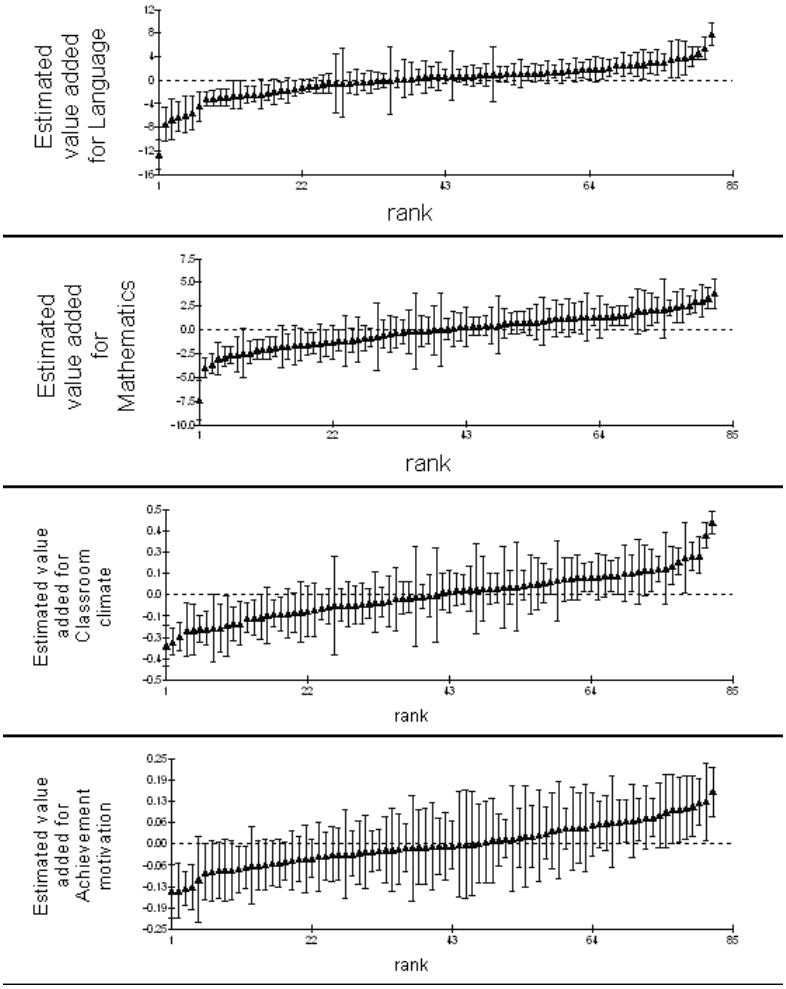
Table 4.6.

Bivariate correlations between prior and final achievement scores

Outcome variable	Biviarate correlation	N
Language	.53**	9,865
Mathematics	.64**	9,660
Classroom climate	.32**	9,164
Achievement motivation	.37**	8,956

Figure 1.

Graphical representations of the estimated value added of school for the cognitive and non-cognitive outcomes



For the dependent variable language higher achievement is associated with high prior achievement scores on language, gender, more affluent family background, Dutch ethnicity, Dutch as only language and higher school types. The results for the dependent variable mathematics show that high performance on the third year test is associated with high scores on the prior achievement test for mathematics, being a Dutch student and higher school types. For the perceived classroom climate, girls score a little higher than boys, students from more affluent backgrounds score slightly higher. Furthermore, the perceived classroom climate is better in the higher school types. However, the perceived classroom climate of the reference group pre-vocation education basic track with additional support, which can be considered as the lowest school type, appears somewhat higher than the perceived classroom climate in combined pre-vocation education middle and basic track and pre-vocational education basic track. With respect to outcome achievement motivation, girls, students from more affluent socio-economic families, non-Dutch students and students speaking only a different language are more motivated to perform better. No significant differences were found between students from different school types after controlling for prior achievement and other student characteristics.

After including predictor variables in the model to estimate value added, the variance accounted for by the school level decreases to 15% for language and 8% for mathematics. Secondary schools appear to have a value added effect on pupils' cognitive achievement in the third grade. A substantial amount of variation (9%) is found at the school level for the non-cognitive outcome classroom climate. Schools differ considerably in how their students perceive the classroom climate, after corrections for differences in student populations. This is not surprising considering the possibilities of schools and teachers to influence behaviour of students within school time. The variance between schools is very small for the achievement motivation (3%).

Differences in effectiveness between schools on the different outcomes, as described above, are more graphically represented in figure 1. In this plot, each triangle depicts the estimated value added for a secondary school. Connected to this triangle is the 95% confidence interval of the estimated value added. On the horizontal axis the secondary schools are ranked from the least to the most effective school. If the confidence interval around the estimated value added includes zero, this school cannot be distinguished from average. A secondary school with a confidence interval above zero can be identified as over performing. And a school with a confidence interval below zero can be identified as underperforming. It appears from

figure one that for language, mathematics and classroom climate a reasonable number of schools can be identified as over- or underperforming. However, for achievement motivation and indicator of value added becomes undiscerning. Only a very small number of schools appear over- or underperforming on the outcome achievement motivation.

4.3.1 Comparing schools on de the cognitive and non-cognitive domain

However, more important in the context of educational accountability is whether school effects are consistent within and between the cognitive and non-cognitive domain. In other words, are schools that perform well on language or mathematics also schools with good classroom climate and schools with motivated students? Both correlations and partial correlations between the different outcomes are presented in table 4.7. Raw correlations among the dependent variables on the school level can be seen as correlations between gross school effects. These correlations were derived from the empty multilevel model. Partial correlations on the school level can be seen as the correlations between the value added estimates of secondary schools. These were derived from the model in which prior achievement and other covariates were included.

Table 4.7.
School level correlation and partial correlation matrix for cognitive and non-cognitive outcomes

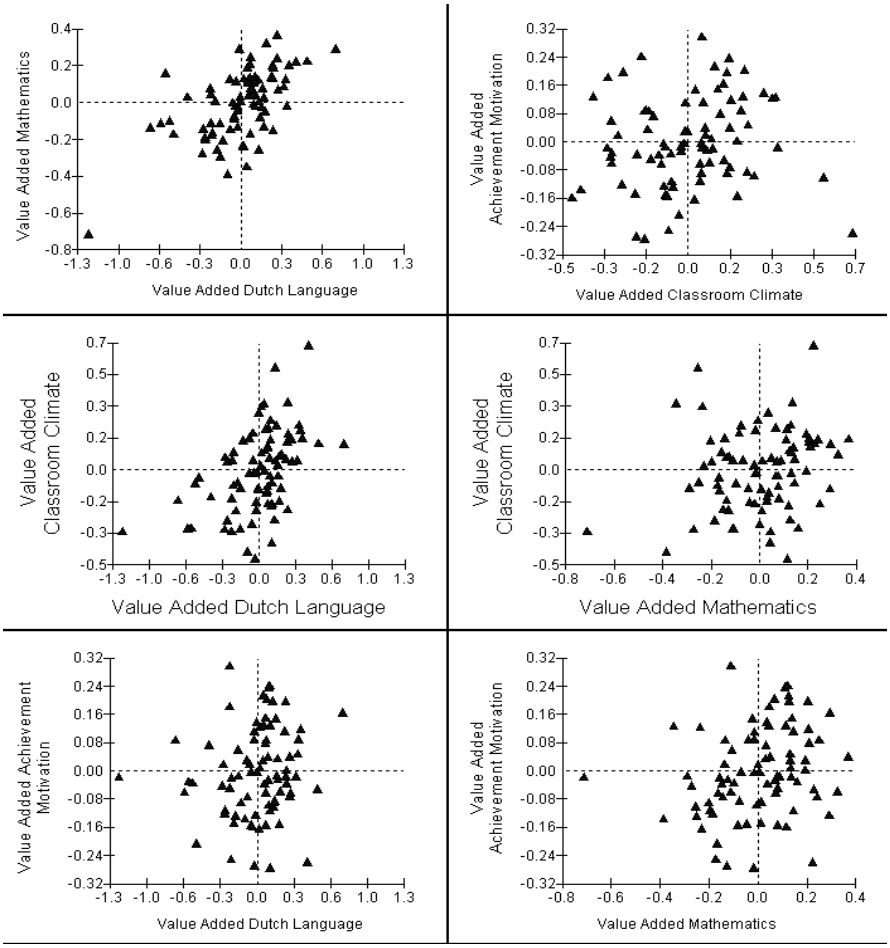
Correlations	**Language**	**Mathematics**	**Classroom climate**	**Achievement motivation**
Language	1			
Mathematics	.91*	1		
Classroom climate	.74*	.63*	1	
Achievement motivation	.12	.20	-.11	1
Partial correlations	**Language**	**Mathematics**	**Classroom climate**	**Achievement motivation**
Language	1			
Mathematics	.60*	1		
Classroom climate	.48*	.21	1	
Achievement motivation	.13	.24*	.10	1

Number of schools 82, * Significant at α=.05 (two tailed)

It is apparent from Table 4.7 that the association between the estimates of gross school effects is larger than the association between the value added estimates. For the gross school effects strong association was found between language, mathematics and the classroom climate. The correlations between the gross school effects of achievement motivation and the other outcomes of education are small and not statistically significant. This means that at the school level achievement motivation is not associated with performance on language, mathematics and classroom climate.

Figure 4.2.

Scatterplots of the association between estimated value added scores of schools for language, mathematics, classroom climate and achievement motivation



After controlling for differences at entry, a moderate correlation of .60 was found between the estimated value added of language and mathematics. Similar correlations between subjects in the cognitive domain in estimated value added of secondary school were found in British and Dutch secondary education (Thomas et al., 1997b; Luyten, 1998). The association between the value added estimates for language and mathematics is presented in a scatterplot in figure 4.2. The moderate positive correlation indicates that schools that perform well on language tend to perform well on mathematics, although the relation is far from perfect. The estimated value added of secondary schools for classroom climate shows a moderate correlation with the estimated value added of language. In other words, schools with positive perceived classroom climate tend to perform well on language. The association between the estimated value added of secondary schools for language and classroom climate are also presented in figure 2. The results of consistency of value added over cognitive and non-cognitive outcomes from this study are somewhat similar to previous research (Thomas et al., 2000; Gray, 2004b). The value added on achievement motivation shows small and non-significant association with the other value added estimates. A similar pattern was found for the gross school effects of achievement motivation.

4.4 Conclusion and discussion

The main focus of this study is the estimation of value added based on multiple outcomes of education, including both cognitive and non-cognitive outcomes. This was done in order to investigate the uni- or multidimensionality of educational effectiveness within the context of educational accountability. The estimated value added of secondary schools for two cognitive and two non-cognitive outcomes were estimated simultaneously in a multivariate multilevel model. Similar to previous studies on schools effects for non-cognitive outcomes, it appears that the variance between schools is considerably smaller for the non-cognitive outcomes. This finding can be explained through the fact that the cognitive domain is explicitly taught in schools. Non-cognitive outcomes are usually a more implicit part of a schools' curriculum (Dijkstra, Karsten, Veenstra, & Visscher, 2001; Peschar, 2004). Nonetheless, the non-cognitive outcome classroom climate shows 8% between school variance. Considering the possibilities of schools and teachers to act on student behaviour within schools and classes, this is not surprising. The very small differences in value added between schools concerning achievement motivation question the possibilities of schools to

influence the motivation of students. Hardly any school can be identified as underperforming on achievement motivation. Because discrimination power of indicators is important for the usefulness in educational accountability, value added based on non-cognitive outcomes of education, which show very little variation between schools, seems not suitable.

The correlations between the estimated value added of secondary schools show somewhat inconsistent results. The strongest association was found between the cognitive outcomes, Dutch language and Mathematics. The moderate positive correlation indicates that schools that perform well on language tend to perform well on mathematics, although the relation is far from perfect. Furthermore, a moderately positive correlation was found between value added of language and classroom climate. It has been argued that educational accountability based on performance indicators on the achievement or progress of students in the cognitive domain might lead to strategic behaviour of schools, such as teaching to the test or a particular focus on the subjects of the test at the expense of other outcomes. However, the moderate positive correlations imply that effectiveness of schools in the cognitive domain doesn't necessarily have to be detrimental for outcomes in the non-cognitive domain. The results support in this sense the multidimensionality of educational effectiveness as advocated in previous studies (Thomas et al., 2000; Gray, 2004b). Furthermore, this finding has implications for the use of value added in educational accountability systems. As a result of inconsistent performance of schools over multiple outcomes, general value added indicators based on average grades mask important differences in effectiveness within the schools (Thomas et al., 1997b; Luyten, 2003). And value added indicators of schools based on a single or a few grades will probably result in biased estimates of the effectiveness of the schools. Therefore, separate indicators should be developed for multiple cognitive outcomes, for example subjects, and non-cognitive outcomes, which are considered as important parts of the curriculum. In making choices which non-cognitive outcomes should be adopted in an accountability system one should consider the following issues: whether or not these outcomes are explicitly taught in the curriculum, if these non-cognitive outcomes show differences between schools and possibilities of valid and reliable of measurements of these non-cognitive outcomes. An important drawback of separate indicators for subjects or groups of subject is that sometimes there are very limited numbers of students taking a particular subject or course. As a results of limited number of students taking particular exams, the uncertainty surrounding the estimated value added will increase (Thomas, 1998).

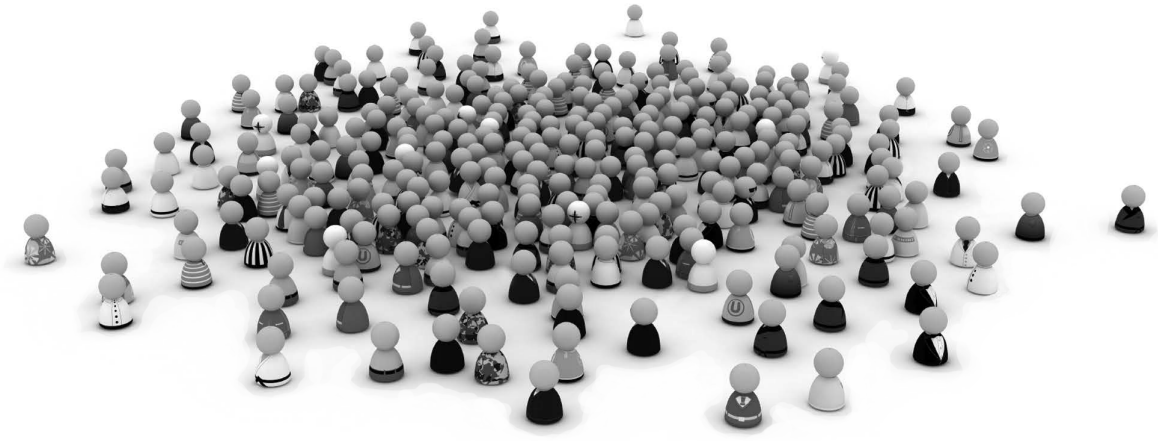
Closely related to the necessity of using multiple indicators for a more detailed and valid view on a schools' effectiveness is the issue of the identification of under achieving schools in educational accountability (Gray, 2004a). This issue arises both from inconsistency in school effects between outcomes, differential school effects for subgroups of students and the stability of school effects over time. Underperformance of a school on a single value added indicator can be established by investigating if the performance of students in a school is significantly lower than the performance of 'similar' students in 'similar' schools, after controlling for prior achievement and other differences at entry. However, in case of multiple indicators of school performance it remains questionable how the results of the indicators can be combined to speak of under-achieving schools.

A number of limitations of the data and the models applied in this study should be considered when interpreting the results. In the first place, only two non-cognitive outcomes were used in this study. The two non-cognitive outcomes certainly must not be looked upon as an optimal or a definitive set and is limited in the coverage of the non-cognitive domain. It is a matter of debate what the non-cognitive outcomes are for which schools should be held (mainly) responsible (Van Damme et al., 2006). Further research in educational effectiveness might include a wider coverage of both the cognitive and non-cognitive domain. For the non-cognitive domain both social and affective outcomes of education might be considered. Citizenship might be an interesting (partly) non-cognitive outcome to consider in future research (Peschar, 2004), because it is by law compulsory for Dutch primary and secondary education since 2006 to actively teach their students to become good citizens. Secondly, attrition analysis revealed some bias on the student level, which might lead to some bias in the estimated value added of individual schools. This would be problematic if such value added indicators would be implemented in educational accountability systems. However, the current study has a strong explorative character in which the aim was to explore the association of value added over multiple indicators.

Furthermore, the models applied only show the association between the estimated value added of secondary schools, while the explanation of the differences in effectiveness within schools between subjects and between domains remains unclear. Many of the known characteristics of effective schools and classes in the cognitive domain appear not significant in explaining differences between classes and schools in the non-cognitive domain (Knuver, 1993). It might be worthwhile to investigate possible explanations for this inconsistency in school effects. These explanations might shed a brighter light on the dimensionality question of educational effectiveness and theory development in Educational Effectiveness Research.

Chapter 5
Value Added Based on Educational Careers in Dutch Secondary Education

This chapter is based on:
Timmermans, A.C., Bosker, R.J., De Wolf, I.F., Doolaard, S., & Van der Werf, M.P.C. (2012). *Value Added Based on Educational Careers in Dutch Secondary Education.* Manuscript submitted for publication.



Abstract

Estimating added value as an indicator of school effectiveness in the context of educational accountability often occurs using test or examination scores of students. This study investigates the possibilities of using scores for educational careers as a complementary indicator. A number of advantages of a value added indicator based on educational careers of students can be formulated, such as: (a) The societal significance of educational position as output measure, (b) the fact that a single indicator can be estimated for an entire school in a differentiated educational system, where not all schools provide education in all tracks. And (c) the expectation that value added based on educational careers leads to other incentives for schools than value added based on test scores. Empirical analysis of Dutch cohort data (VOCL'99) for secondary education showed considerable differences in effectiveness between schools in the careers of students. Furthermore, differential school effects were found for both socio-economic status and prior achievement. The phenomena of differential school effects for socio-economic status and prior achievement are linked to differences between schools in the tracks in which the schools provide education.

5.1 Introduction

In the last decade, most countries have introduced a system of educational accountability. These accountability systems give insight in the educational performance of educational institutions and are used to inform governments, students and parents. Most accountability systems in education use either the percentages of pupils that pass the exams and/or test or examination score indicators to measure the performance of schools. In general, indicators can be divided in two categories, namely performance indicators based on test or examination scores and efficiency indicators based on passing or failing examinations and possible grade retention. Examples of educational accountability systems with a strong focus on the performance or attainment of students are the Dutch, English, Scottish and Belgian educational accountability systems.

In recent years, value added indicators have been adopted in many educational accountability systems. Value added indicators were developed to make a fair comparison of the performance of educational institutions (Meyer, 1997). In most of these value added indicators performance of students on tests or examinations is used to estimate differences in performance between educational institutions, while controlling for differences in student intake at entry of a formal stage of schooling (Raudenbush et al., 1986; Aitkin et al., 1986; Raudenbush et al., 1995; Goldstein, 1997; Bosker et al., 2001). These value added indicators based on test or examination scores are usually interpreted as the difference in test performance of students in school J and the average school for students with a comparable level of prior achievement (and possible other student characteristics) (Willms et al., 1989; Raudenbush et al., 1995). Examples of well-known value added models used in educational accountability systems are Contextualized Value Added (Ray, 2006; Ofsted, 2010), the Tennessee Value Added Assessment System (Sanders et al., 1994; Sanders, 2003) and the Colorado growth curve model (Betebenner, 2007; Betebenner, 2009). These value added indicators are examples of performance indicators, as they are based on test or examination scores. Many studies have questioned the validity of value added indicators based on test or examination scores in high stakes educational accountability systems (McCaffrey et al., 2003; Cantrell et al., 2007; Koretz, 2008; Kane et al., 2008; Rothstein, 2008). Their main critiques concern the possibilities of schools to perform strategic behaviour like teaching for the test and test manipulation.

Moreover indicators based on test scores have limitations in differentiated educational systems in which different tracks use different tests or examinations. Especially in countries in which schools offer different tracks this is an important disadvantage for the usefulness of performance indicators in educational accountability, because it's not possible to compare the value added of entire schools. As governmental bodies with accountability tasks attempt to formulate transparent and simple, though valid frameworks for educational accountability, they might want indicators of the performance of an entire school.

An alternative for using test or examination scores in the estimation of value added indicators that might overcome these disadvantages is using the educational position or the stage of educational career of students. This latter can be seen as an efficiency indicator. In that case the value added indicator can be used to compare schools on how well they guide students in reaching the optimal grade and track, given their intake. Value added based on the educational position of students can be interpreted as the difference between educational careers or educational opportunities of students in school J and the average school for students with a comparable level of prior achievement (and possible other student characteristics). Therefore, the interpretation of both value added indicators is rather similar.

In this article we will discuss the differences between value added indicators based on test scores and a value added indicator based on the position of the educational career. Differences between both types of indicators will be discussed on several aspects of validity, reliability, transparency and possibilities for strategic behaviour. Given that these aspects are important criteria for the usefulness of a performance indicator in educational accountability. Special attention is given to some possibilities of value added indicators based on educational careers in differentiated educational systems, such as many educational systems in Europe. Examples of such differentiated educational systems are the Dutch, Belgian, German, Poland, Russian and Irish educational systems. Cohort data from Dutch secondary education will be used in this article as an example of modelling value added on educational careers of students in a differentiated educational system. Because of the way tracks and grades are ordered in Dutch secondary education, it is possible to construct one variable that indicates the position of a student in the system– at the so-called 'educational ladder' - at every point in time. A detailed description of Dutch secondary education and the construction of this 'educational ladder', used in the empirical part of this research, can be found in the methods section of this study. In general, students get a higher score on the educational ladder as they reach higher grades and/or higher tracks.

5.1.1 Comparing value added models on aspects of validity, reliability, transparency and strategic behaviour

Indicators based on educational position might differ from indicators based on test scores with respect to the societal significance of output measure on which the performance indicator is based. This societal significance lies in the fact that each track offers different access to further education. The position of students within the educational system, therefore, partly determines the future educational opportunities and subsequent job opportunities for students. This societal significance of an outcome measure for stakeholders might influence their perceived value of the performance indicator, so an indicator based on the educational position could be more relevant to the users than an indicator based on test scores.

A considerable number of articles have raised concerns about the reliability and validity of value added indicators. These concerns regard value added indicators based on test or examination scores. For example, ignoring student mobility might lead to bias and underestimation of the estimated value added indicator (Goldstein et al., 2007; Leckie, 2009; Timmermans, Snijders, & Bosker, 2012). Moreover, using average test scores over multiple subjects might mask important differences in effectiveness between departments within schools (Luyten, 2003). Furthermore, measurement error in the control variables leads to an underestimation of the estimated value added for schools (Hill, & Rowe, 1996). For several of the validity issues, such as bias through student mobility (Goldstein et al., 2007; Leckie, 2009), ceiling effects (Schagen, 2006; OECD, 2008) or measurement error (Woodhouse et al., 1996; Goldstein et al., 2008), statistical solutions exists at the expense of the transparency of the value added indicators. These validity threats have not yet been investigated for value added based on the educational careers of students and therefore the impact of these threats on the indicator is so far unknown. Similar validity threats might be expected as the methodology for estimating value added based on educational careers is fairly similar to traditional value added measures. However, the extent of the validity threats depend on a large number of factors, such as the between school differences, the amount of between school mobility, the within school mobility and the covered time span.

Differences between value added models based on test scores and educational careers arise with respect to the possibilities of strategic behaviour of schools. As Campbell's law says; "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." (Campbell, 1976) This law is regularly associated with high stakes testing in education

and the use of performance indicators in educational accountability. This phenomenon is also known as strategic behaviour (De Wolf & Janssens, 2007). In the next paragraphs the differences between both value added indicators will be discussed for the different types of strategic behaviour.

A substantial body of literature describes increasing placement of students in special education, consequences for students with special educational needs and sickness at the test date as examples of reshaping the test pool in educational accountability systems (Figlio et al., 2002; Jacob, 2005; Nichols & Berliner, 2005; Cullen et al., 2006; Lemke et al., 2006; Swanborn et al., 2008; Jones, 2008). The Dutch secondary education system is characterized by a relatively large number of possibilities for track mobility compared to several other European tracked educational systems, such as the German tracked system (Jacob & Tieben, 2007). Intermediate downward mobility might be used by schools to inflate their examination scores and thereby to artificially enhance their value added. After all, a weak or struggling pre-university student might be a very good higher general secondary education student. Intermediate downward mobility for these students increases the average examination scores for both the pre-university and higher general secondary education track. This strategy might increase examination scores, but not the position students hold within an educational system. Value added based on the position within the educational system provides schools with opposite incentives for strategic behaviour by reshaping the test pool. Grade retention or intermediate downward mobility between tracks leads to lower scores for these students on the educational ladder, the variable on which we can map the score for the educational career. Using indicators based on educational careers of students in educational accountability might give schools incentives to places as many students as possible in the higher tracks, possibly resulting in lower scores for these students on national examinations. However, the reasons for up- or downward mobility cannot be derived from large scale administration data.

A second method of strategic behaviour is known as teaching to the test. Several studies have shown that high stakes testing and educational accountability have forced teachers and schools to align their curriculum to the areas tested (Jacob, 2005; Nichols et al., 2005; Koretz, 2005; Jones, 2008). This form of strategic behaviour is associated regularly with standardized high stakes testing with a large focus on mathematics and reading skills. Teaching to the test or reallocating recourses remains an option to a certain extent if the educational careers of students are considered as outcome measure. For example, in secondary education schools might discourage students in taking extra subjects alongside compulsory subjects.

Test manipulation is the third method of strategic behaviour. Test manipulation includes practices such as cheating, offering students additional resources while making the tests or additional instruction during the test. Studies in both the United States and the Netherlands have shown that some teachers apply test manipulation to increase test scores (Jacob et al., 2003; Jacob, 2005; Nichols et al., 2005; Swanborn et al., 2008; Amrein-Beardsley, Berliner, & Rideau, 2010). Since the position of students on the educational ladder can be derived through administrative data, test manipulation cannot be applied to artificially increase value added based on the educational ladder. However, the reliability of the data, whether it is administration data or test scores, remains an important consideration for the validity and usefulness of value added indicators.

To summarize, a number of supposed advantages of a value added indicator based on educational careers of students can be formulated, such as: (a) the societal significance of educational position as output measure, (b) the fact that a single indicator can be estimated for an entire school in a differentiated educational system, where not all schools provide education in all tracks. And (c) the expectation that value added based on educational careers leads to other incentives for schools than value added based on test scores. It might be considered to use value added based on educational careers and test or examination scores as complementary indicators in educational accountability, because both indicators provide valuable information concerning the effectiveness of schools and tracks despite several flaws. Given the opposite incentive, using multiple indicators might be beneficial for the robustness of an accountability system with respect to strategic behaviour (Koretz, 2003).

Despite the supposed advantages of value added indicators based on educational positions of students, it is important to assess the possibilities for estimating such an indicator and to assess the validity of the indicator. The empirical analysis focusses on estimating value added based on educational careers and two aspects of the validity of value added indicators. The first validity aspect of value added is differential school effects, which refers to differences in effect of school for particular subgroups of students (Nuttall et al., 1989; Sammons et al., 1993; Thomas et al., 1997a; Veenstra, 1999; Gray et al., 2004), for example based on prior achievement, gender or ethnic background. In case of differential school effects, a single value added estimate for all subgroups of students within schools might mask differences in effectiveness within schools. The second validity aspect refers to one of the supposed advantages of value added based on educational positions. One of the supposed advantages is that value added based on educational positions can be estimated for entire schools within differentiated educational systems. To test whether this advantage holds we examine

to what extent the school composition in terms of tracks can account for differences in value added based on educational careers. Not all schools provide education in all school tracks, in this final analysis we test whether schools with different structures can be fairly compared.

The following research questions will be answered in this study.

1. Can a value added indicator based on the educational ladder distinguish secondary schools in terms of their overall effectiveness?
2. Are schools differentially effective for specific subgroups of students?
3. To what extent can the school composition in terms of tracks explain differences in value added based on educational careers?

The second section of this study gives an overview of the data and methods used. In the third section, the empirical results will be presented. Implications of the results for both educational effectiveness research and educational accountability will be discussed in the last section of this paper.

5.2 Method

5.2.1 Sample

The data used in this study were collected as part of a national longitudinal study in secondary education in the Netherlands, the "Cohort Studies in Secondary Education" (Dutch abbreviation: VOCL). The data concerned students who entered the first grade of Dutch secondary education in the Netherlands (comparable to the 7th grade in the United States) in the year 1999, also called the VOCL'99 cohort. The original two-stage sample involved 108 secondary schools and 19,391 students. This sample has been considered as representative of the schools and students in the Dutch secondary education (Kuyper et al., 2003b).

Data collection of the VOCL'99 started at the moment of secondary school entry in 1999, when the students were about 12 years old. Students' prior achievement level was assessed with a test during the first year of secondary education. Every year, information about students' grade and school track was collected. We used this information to define students' educational career position. A demographical sketch of the participants in this study is reported in Table 5.1, immediately after the description of the variables.

A selection of students and schools was made for the analysis based on the following criteria. For students, complete records had to be available for all covariates. On the school level, units with less than 36 students with complete records were excluded from the analysis, because for a reliability of .80, given an intraclass correlation (ICC) of .10, at least 36 lower level units should be available. In total 8,635 students in 67 schools were selected for the current study. Further below, the issue of potential attrition bias is addressed.

5.2.1.1 Dutch secondary education

Students in Dutch secondary education are placed in a specific track based on their scholastic aptitude around the age of 12. In total there are five ordered track-levels in Dutch secondary education. The duration of the tracks varies between four (the three lowest tracks) and six years (the highest track). Each track offers different access to further education and the final examinations differ between the tracks in level and content. The pre-university track (the highest track) is the only one that directly prepares students for university education. Higher general secondary education is the second highest track and prepares the student for further education in higher vocational education or universities for applied sciences. The three pre-vocational education tracks prepare the students for further education in senior secondary vocational education, although these pre-vocational education tracks differ in level and further educational opportunities. Students from the pre-vocational basic track may enter training programmes in senior secondary education at the basic level. Pre-vocational education middle track students may enter training programmes in senior secondary education at the level of professional training and students from the pre-vocational theoretical track may enter training programmes in senior secondary education at the middle management level. Grade repetition within tracks and intermediate upward or downward mobility between the tracks is possible, as students can change tracks depending on their grades. Furthermore, after successfully completing one of the tracks students gain access to further education in the next higher track level. For example, a student who successfully finished higher general secondary education gains access to the fifth year of pre-university education.

5.2.2 Variables

Of focal interest in this study is the criterion variable *students' educational career*. *Prior achievement*, *socio-economic status* and *ethnicity* functioned as covariates in the analyses

where differences in effectiveness between schools are established in how well they guide students in reaching the optimal grade and track. The variables and their instrumentation used in the analysis are discussed below. Distributional characteristics of students and schools are presented in Table 5.1 and 5.2.

Score on the educational ladder. This measure was originally developed to map the grade and track of a student within a differentiated educational system (Bosker & Van der Velden, 1985; Bosker & Van der Velden, 1989). The educational ladder used for the current study is presented in Figure 5.1. This educational ladder differs slightly from the original, due to changes in the educational system (Claassen & Mulder, 2003; Driessen, 2011; De Boer, 2009; Roeleveld, Driessen, Ledoux, Cuppen, & Meijer, 2011).

Figure 5.1
Educational ladder in Dutch secondary education

Score on the educational ladder	Pre-vocational education basic track	Pre-vocational education middle track	Pre-vocational education theoretical track	Higher general secondary education	Pre-university education
12					**diploma**
11					6 years
10				**diploma**	5 years
9				5 years	4 years
8			**diploma**	4 years	3 years
7		**diploma**	4 years	3 years	2 years
6	**diploma**	4 years	3 years	2 years	1 year
5	4 years	3 years	2 years	1 year	
4	3 years	2 years	1 year		
3	2 years	1 year			
2	1 year				

The scores on the Educational Ladder range from 1 to 12. Completion of the highest track is valued at 12 points and starting secondary education at the lowest track is valued at 2 points (1 point is for starting special needs education). Thus, a high score reflects a high position of a students' educational career, and a low score reflects a lower position on a students' educational career. The score on the education ladder increases by 1 point when a student proceeds to the next year or when he or she moves one track higher but stays in the same grade. A student's score on the

education ladder is obtained by subtracting the number of years that a student needs to go to get to the top (i.e., direct access to university education) from the maximum score of 12 points. For example, a student's score in the 6th year and thus final grade of the highest track is 11, a student's score in the 5th year is 10, a student's score in the 5th year of one track lower is 9, and so on.

For the current study we established the highest position of the educational career of the students on the educational ladder after the fourth year in secondary education. The highest position after four years is established, because the shortest vocational tracks in secondary educational take four years and therefore it allows a fair comparison between schools that provide education in different tracks. Dropouts are also included in the analysis and these students will get the highest position they reached on the educational ladder before they dropped out. For the students in the highest track we assessed whether these students were promoted to the fifth grade and for the students in the lower three tracks we assessed whether they graduated or not. In the VOCL'99 cohort, students with special educational needs (lwoo) are assigned to the pre-vocational education basic track. After four years, the students' scores on the educational ladder vary between 3 and 11 points.

Prior achievement. Students' prior achievement was assessed at the end of primary school by administering the Cito Eindtoets, a test 80% of the Dutch pupils take when they are halfway 6th grade (at the end of their primary school). Scores on this test were retrieved form the primary schools. This Cito-test was developed by the Dutch National Institute for Educational Measurement (CITO), which is the Dutch equivalent of the Educational Testing Service in the United States. The test has been designed to provide teachers and students an objective measure of students' achievement level, and to support teachers' recommendations and students' choices for secondary school track types. The test consists of 200 multiple-choice items, divided over three subtests: Dutch Language, Arithmetic, and Study Skills. Test scores ranged from 505 to 550. The reliability of the test (*KR20*) is .95.

Socio-economic status. SES was measured by the highest educational level completed by one or both of the student's parents. This variable consisted of six categories, ranging from only primary education to post-graduate. In the analysis socio-economic status was used as a continuous variable.

Ethnicity. Information about the ethnic origin of students was gathered by asking the parents in which country they were born. Students' ethnicity was operationalized as a dichotomous variable with the categories indigenous (coded as 0) and minority (coded as 1). Only if both parents and the student were born in the Netherlands was

the student considered to be indigenous; in all other cases, the student was considered to be a minority student (Kuyper et al., 2003a).

Table 5.1
Distributional characteristics on the student level

Characteristic	Min.	Max.	*M.*	*SD.*	%
Educational career	3.00	11.00	7.88	1.60	
Prior achievement	505.00	550.00	536.35	9.02	
Socio-economic status	2.00	7.00	4.06	1.12	
Ethnicity (minority students)					17.3

N=8,635

Table 5.2
Distributional characteristics on the school level

Characteristic	Min.	Max.	*M.*	*SD.*	%
Prior achievement school composition	519.84	547.33	535.23	6.58	
Socio-economic status school composition	2.82	5.47	4.00	0.51	
School composition in tracks					
Only pre-university education					7.5
Pre-university education, higher general secondary education (and pre-vocational education theoretical track)					19.4
Only pre-vocational education theoretical track					7.5
All pre-vocational education tracks					7.5
All tracks					58.2

N=67

School composition for prior achievement and socio economic status Both prior achievement and socio-economic status were aggregated to the school level to assess whether the average prior achievement and the average socio-economic status of students within schools have an impact on the students' educational career. In this aggregation process all available data on these variables are considered, which implies that composition variables were based on all observed students within schools for these particular variables. This includes information from students that were excluded from the analysis due to missing values on other variables.

School composition in terms of provided tracks Four categories of school composition in terms of tracks have been defined. Namely, (a) schools that only provide education in the pre-university track, (b) schools that provide education in the general tracks (pre-university, higher general secondary and pre-vocational theoretical track), (c) schools that only provide education in the pre-vocational education theoretical track, (d) school that provide education in multiple pre-vocational education tracks and (e) schools that provide education in all tracks.

5.2.3 Attrition

Due to missing values on one or more covariates and selection criteria on the school level, 10,756 students and 41 schools of the original sample were lost from the analysis. Prior achievement is the variable with the largest amount of missing values (8,821 missing). Comparing students who were included in the analysis and the students who were excluded from the analysis revealed some possible sources of attrition bias. Students included in the analysis reach 0.50 points higher on the educational ladder than excluded students (t=-20.85; df= 19072; p<0.001). Students included in the analysis performed on average higher at the prior achievement test (t=-13.63; df=2743; p<0.001). Furthermore, a larger proportion of the group students included in the analysis is Dutch (χ^2=39.39; df=1; p<0.001) compared to the excluded students. For socio-economic status we did not find signs of attrition bias on the student level. On the school level, no signs of attrition were found for the number of students within the school and for the school composition variables of prior achievement and socio-economic status.

5.2.4 Methods of analysis

Based on the idea of the educational ladder, the value added of a school can be estimated for the educational careers of the students with similar methods as traditional value added models based on test scores. This means that we can use (multilevel) regression models in which prior achievement and other student characteristics at entry of a formal stage of education are included as covariates. Hierarchical linear models, or multilevel models, were used for estimating school differences in students' educational careers after four years of secondary education, using MLwiN version 2.24 software (Rasbash et al., 2009). Hierarchical linear models are considered the most appropriate to estimate the effects of schools because they take the hierarchical structure of the data into account(Snijders et al., 1999).

The position of students on the educational ladder is used as the dependent variable in the analysis. To control for differences in student intake of schools at entry we use the student level variables prior achievement, socio-economic status, second language and ethnicity, as well as the school composition variables average prior achievement and average socio-economic status as control variables. The first model (Model 0) is an empty model or a gross school effects model. In model 1, prior achievement is included in the model as control variable. This model can be seen as a very simple value added model. Thereafter in model 2 other student level covariates and school composition variables were included in the analysis. In the analysis the continuous control variables socio-economic status and prior achievement and the composition variables at the school level were centred around their grand mean.

In previous research, the scores of students on the educational ladder have been used as interval level variable. However, this implies equal distances between the different school tracks (1 point on the ladder) and implies that this distance between school tracks is equal to one year of education. It is questionable whether the differences between school tracks are all equal and comparable to one year of education. To test whether or not the educational ladder can be analysed on as interval variable, the fit of an interval and an ordinal model have been compared. The fit of an empty ordinal multilevel model (DIC= 64615.5) appeared substantially better than for an empty linear multilevel model (DIC= 68559.0), which implies that at best this outcome variable should be analyzed on the ordinal level.

For the analysis of the ordinal dependent variable a two-level multinomial ordered logit model was employed using the MCMC algorithms in MLwiN (Rasbash et al., 2009; Browne, 2009). This method of analysis is also known as the multilevel ordered logistic regression model or multilevel proportional odds model (Hedeker, 2008). For estimating the parameters MCMC algorithms were preferred to quasi-likelihood methods because they yield less biased estimates in the multilevel logistic regression analysis, especially in the case of estimating the random-effects variance (Browne & Draper, 2006). This is important for the analysis of value added where the focus lies on the variance in residuals on the level of secondary schools. In this analysis score 11 (the highest) on the educational ladder was used as reference group. However, using the highest group as point of reference implies that a negative coefficient of the control variables indicates a positive association and vice versa. We used the Deviance Information Criterion (DIC), a combined measure of model fit, and model complexity to compare the model fit of the models estimated on the basis of the MCMC algorithms (Spiegelhalter, Best, Carlin, & Van der Linde, 2002). Models with smaller DIC values are to be preferred to models with larger DIC values. A

difference of 5 points between the models is considered as a substantial improvement of model fit.

5.3 Results

5.3.1 Differences among schools in educational careers of students

Results of the ordered logistic multilevel regression analyses are presented in Table 5.3. It is apparent from Table 5.3 that 34.5% of the variance in educational careers in Model 0 is accounted for by secondary schools. However, no control variables were included in model 0, and therefore, results of this model can be seen as gross school effects instead of value added estimates. Furthermore, in Model 1, 9.8% of the variance in educational careers can be accounted for by secondary schools, after controlling for differences in prior achievement of the students. And in Model 2, if other student background characteristics and school composition variables were included in the analysis, only 7.1% of the total variance remains accounted for by the school level.

The category "educational ladder score 11" is the reference group in the models. Therefore, a positive coefficient in the table implies an increase in the probability of obtaining lower scores on the educational ladder and a decrease in the probability of achieving higher scores on the educational ladder. A negative coefficient in the model implies an increase in the probability of achieving higher scores on the educational ladder. From model 1 it can be seen that students with high prior achievement scores tend to reach higher scores on the Educational Ladder. Besides the effects of prior achievement, significant effects were found for the student's socio-economic status as a predictor of a student score on the Educational Ladder in Model 2. Students from more affluent families tend to reach higher scores on the Educational Ladder. After including prior achievement and socio-economic status, no significant differences were found between Dutch and minority students.

The estimated value added scores of secondary schools derived from Model 1 are presented in Figure 5.2. In this figure, the secondary schools are ranked from the most to the least effective in terms of their added value on the educational careers of their students. Similar to the coefficients in the models, negative residuals on the log odds scale at the school level imply an increase of the probability of achieving higher scores on the educational ladder. Each triangle represents the estimated value added of a secondary school, surrounded by its 95% confidence interval. From Figure 5.2, it is apparent that 18 secondary schools (26.9%) can be identified as effective or

overperforming, as these schools have negative estimated value added on the log odds scale and their confidence interval does not include zero. These schools reach higher scores on the Educational Ladder than might be expected from their students given their prior achievement. Furthermore, 17 schools can be identified as ineffective or underperforming (25.4%) and 32 secondary schools (47.8%) can be identified as average. The ineffective or underperforming schools reach lower scores on the educational ladder than might be expected from their students given their prior achievement. From Figure 5.2 it becomes apparent that the value added indicator based on the educational ladder discriminates between average, over- and underperforming schools.

Figure 5.2
Estimated value added of secondary schools using educational careers of their students; based on the results of Model 1.

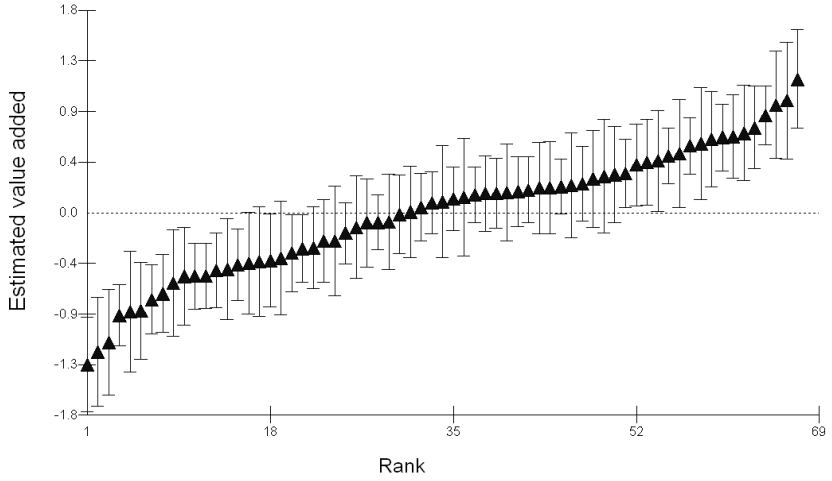


Table 5.3

Results of the MCMC estimation of the two level multinomial ordered logistic value added model based on the educational ladder

	Model 0			Model 1			Model 2		
	Coefficient	**S.E.**	**Coverage Interval (95%)**	**Coefficient**	**S.E.**	**Coverage Interval (95%)**	**Coefficient**	**S.E.**	**Coverage Interval (95%)**
Fixed Part									
Educational ladder score 3 (β_1)	-5.30*	0.20	[-5.69;-4.92]	-6.56*	0.15	[-6.85;-6.26]	-6.70*	0.14	[-6.99;-6.42]
Educational ladder score 4 (β_2)	-4.29*	0.17	[-4.64;-3.96]	-5.47*	0.11	[-5.69;-5.24]	-5.61*	0.11	[-5.82;-5.40]
Educational ladder score 5 (β_3)	-2.94*	0.16	[-3.27;-2.64]	-3.95*	0.09	[-4.13;-3.76]	-4.07*	0.08	[-4.23;-3.91]
Educational ladder score 6 (β_4)	-1.45*	0.16	[-1.77;-1.15]	-2.07*	0.09	[-2.23;-1.89]	-2.16*	0.07	[-2.30;-2.02]
Educational ladder score 7 (β_5)	-0.82*	0.16	[-1.14;-0.53]	-1.22*	0.08	[-1.37;-1.04]	-1.29*	0.07	[-1.43;-1.15]
Educational ladder score 8 (β_6)	0.74*	0.16	[0.42;1.03]	0.90*	0.08	[0.74;1.07]	0.87*	0.07	[0.74;1.01]
Educational ladder score 9 (β_7)	1.87*	0.16	[1.55;2.16]	2.34*	0.09	[2.18;2.52]	2.34*	0.07	[2.21;2.49]
Educational ladder score 10 (β_8)	10.37*	1.27	[8.46;13.40]	11.12*	1.30	[9.21;14.28]	11.19*	1.28	[9.27;14.27]
Prior achievement (β_9)				-0.19*	0.003	[-0.20;-0.18]	-0.18*	0.003	[-0.19;-0.18]
Socio-economic status (β_{10})							-0.36*	0.02	[-0.37;-0.30]
Minority (β_{11})							-0.04	0.06	[-0.15;0.08]
Average socio-economic status (β_{12})							-0.33	0.23	[-0.77;0.15]
Average prior achievement (β_{13})							-0.02	0.02	[-0.06;0.02]
Random Part									
Level 2 random effects									
σ^2_{u0} (School level intercept variance)	1.73	0.3	[1.21;2.46]	0.36	0.07	[0.24;0.53]	0.25	0.05	[0.16;0.37]
ICC			.345			0.098			.071

* p<.05 (two-tailed) Coefficients are reported on the log odds scale.

Table 5.3 (Continued)

Results of the MCMC estimation of the two level multinomial ordered logistic value added model based on the educational ladder

Model fit	Model 0	Model 1	Model 2
DIC	27709.99	24010.3	23747.3
pD	72.7	68.6	68.7
Units level 2: Schools	67	67	67
Units level 1: Students	8,635	8635	8635

Table 5.4

Differences between secondary schools in educational careers of their students presented through model-based cumulative predicted probabilities for students with average prior achievement (Model 1)

	Predicted probabilities on educational ladder score*							
	4	5	6	7	8	9	10	11
Very effective (-1 log odds)	.999	.999	.993	.956	.902	.542	.207	.001
Effective (-0.5 log odds)	.999	.997	.988	.929	.848	.401	.137	.001
Average (0 log odds)	.999	.996	.981	.888	.771	.288	.088	.001
Ineffective (+0,5 log odds)	.998	.993	.969	.828	.672	.197	.055	.001
Very ineffective (+1 log odds)	.996	.989	.950	.745	.554	.130	.034	.001

* The predicted probabilities for a score 3 on the educational ladder are excluded from the table since the cumulative probabilities are equal to one for all schools.

To get some indications of the size of the differences between effective and ineffective schools in the positions of their students on the Educational Ladder after four years of education, model based cumulative predicted probabilities are calculated for the effective, ineffective and the average school for a student with average prior achievement scores. These cumulative predicted probabilities are presented in Table 5.4. Based on the cumulative probabilities it appears that an average student in a very effective school has a much larger probability of reaching the higher scores on the educational ladder, as the cumulative probability of reaching a score of eight or higher on the educational ladder is .902 (diploma in pre-vocational education theoretical track or higher).This average student has a probability of (1-.902) of reaching a score of seven or lower. While a similar student in a very ineffective school has a cumulative predicted probability of .554 of reaching a score of eight or higher on the educational ladder. The chance of reaching a score seven or lower in an ineffective school for this average student is .446 (1-.554). This is a difference of .348 (or almost 35%) in the probability of reaching a score on level eight or higher between the very effective and very ineffective school for an average student.

5.3.2 Differential effectiveness of schools in educational careers

To assess whether secondary schools are equally effective in promoting the educational careers of all students a random slopes model was estimated. The results obtained from this random slopes model (Model 3) are presented in Table 5.5. The model with random slopes for prior achievement and socio-economic status (Model 3, DIC=23328.8) showed an improved model fit compared to Model 2 (DIC=23747.3), as presented in Table 5.3. This improved model fit indicates the existence of differential schools effects.

The random slopes for prior achievement indicate differences between secondary schools in the relationship between prior achievement and the score on the Educational Ladder. Differences between low and high prior achievement students in scores on the Educational Ladder are larger in schools with a steep slope for prior achievement and smaller in schools with a more flat slope for achievement. The most striking result to emerge from Model 3 is that there appears to be no association (r=.09) between the random slopes of prior achievement and the random intercepts. This means that having a steep or an even slope for prior achievement is not related to a schools average effectiveness.

Model-based predicted probabilities for three typical schools, with an average overall effectiveness and steep, average and flat slopes, are presented in Table 5.6 to

illustrate the random slopes for prior achievement. For each school the predicted cumulative probabilities of students with a low, average and high prior achievement were given. Because grand mean centering for prior achievement was applied in the model, the predicted probabilities of student with average prior achievement are equal for these three schools. The differences between high and low prior achievement students in predicted probabilities are the smallest in schools with a relatively flat slope (b=.11). For example, the probability of reaching a score of eight or higher is .522 for a low prior achievement student and .888 for a high prior achievement student. The difference in probability between a low and high prior achievement student in this school is .366 (.888-.522) or almost 37% for reaching a score of eight or higher. The difference of reaching a score of eight or higher between high and low prior achievement students is .570 (.937-.367) in a school with an average slope. In a school with a steep slope the difference in probability reaches .730 (.966-.236) between low and high prior achievement students in reaching a score of eight or higher on the educational ladder. The differences appear smaller at both ends of the educational ladder, because the score 11 (sixth grade pre-university education) and the lowest scores are uncommon for these students.

Table 5.5

Results of the MCMC estimation of the two level multinomial ordered logistic value added model based on the educational ladder for differential school effects

	Model 3		
	Coefficient	**S.E.**	**Coverage Interval (95%)**
Fixed Part			
Intercept educational ladder score 3 (β_1)	-6.45*	0.15	[-6.75;-6.17]
Intercept educational ladder score 4 (β_2)	-5.36*	0.11	[-5.58;-5.14]
Intercept educational ladder score 5 (β_3)	-3.84*	0.09	[-4.02;-3.66]
Intercept educational ladder score 6 (β_4)	-1.95*	0.08	[-2.11;-1.79]
Intercept educational ladder score 7 (β_5)	-1.08*	0.08	[-1.24;-.093]
Intercept educational ladder score 8 (β_6)	1.18*	0.08	[1.03;1.34]
Intercept educational ladder score 9 (β_7)	2.77*	0.09	[2.59;2.93]
Intercept educational ladder score 10 (β_8)	11.77*	1.26	[9.89;14.79]
Prior achievement (β_9)	-0.18*	0.01	[-0.20;-0.16]
Socio-economic status (β_{10})	-0.31*	0.03	[-0.37;-0.26]
Minority (β_{11})	-0.02	0.06	[-0.13;0.10]
Average socio-economic status (β_{12})	-0.125	0.24	[-0.61;0.35]
Average prior achievement (β_{13})	-0.07*	0.02	[-0.11;-0.03]
Random Part			
Level 2 random effects			
σ^2_{u0} (school level intercept variance)	0.28	0.06	[0.18;0.43]
σ^2_{u9} (slope variance prior achievement)	0.005	0.001	[0.003;0.007]
σ^2_{u10} (slope variance socio-economic status)	0.019	0.007	[0.009;0.035]
$\sigma_{u0,9}$ (covariance intercept & prior achievement)	0.003	0.008	[-0.011;0.019]
$\sigma_{u0,10}$ (covariance intercept & socio-economic status)	-0.05	0.02	[-0.08;-0.02]
$\sigma_{u10,9}$ (covariance prior achievement & socio-economic status)	0.005	0.002	[0.002;0.010]
Model fit			
DIC			23328.8
pD			128.9
Units level 2: Schools			67
Units level 1: Students			9635

* p<.05 (two-tailed) Coefficients are reported on the log odds scale.

Table 5 .6

Differential school effects of secondary schools in educational careers of their students based on prior achievement presented through model-based predicted probabilities (Model 3)

		Predicted probabilities on educational ladder score*							
		4	5	6	7	8	9	10	11
Flat slope	Low prior achievement**	.996	.988	.946	.723	.552	.102	.023	.000
	Average prior achievement	.999	.995	.979	.875	.746	.235	.059	.001
	High prior achievement	.999	.998	.992	.950	.888	.453	.145	.001
Average slope	Low prior achievement	.992	.977	.902	.581	.367	.057	.012	.000
	Average prior achievement	.999	.995	.979	.875	.746	.235	.059	.001
	High prior achievement	1.00	.999	.996	.973	.937	.609	.241	.001
Steep slope	Low prior achievement	.985	.957	.830	.424	.236	.031	.007	.000
	Average prior achievement	.999	.995	.979	.875	.746	.235	.059	.001
	High prior achievement	1.00	1.00	.998	.985	.966	.746	.374	.001

* For Dutch students with an average socio-economic status

** The operationalization of low prior achievement is 1 standard deviation below average, high prior achievement means 1 standard deviation above average

Table 5.7

Differential school effects of secondary schools in educational careers of their students based on socio-economic status presented through model-based predicted probabilities (Model 3)

		Predicted probabilities on educational ladder score***							
		4	5	6	7	8	9	10	11
Flat slope	Low socio-economic status	.999	.996	.972	.882	.759	.247	.063	.001
High intercept	Average socio-economic status	.999	.997	.984	.905	.799	.293	.078	.001
	High socio-economic status	.999	.997	.988	.923	.834	.344	.097	.001
Average slope	Low socio-economic status	.998	.993	.970	.832	.675	.178	.042	.001
Average intercept	Average socio-economic status	.999	.996	.979	.875	.746	.235	.059	.001
	High socio-economic status	.999	.997	.985	.909	.806	.303	.081	.001
Steep slope	Low socio-economic status****	.997	.990	.956	.767	.580	.126	.028	.000
Low intercept	Average socio-economic status	.998	.994	.972	.839	.686	.185	.044	.001
	High socio-economic status	.999	.996	.982	.892	.775	.265	.068	.001

*** For Dutch students with an average prior achievement

**** The operationalization of low socio-economic status is 1 standard deviation below average, high socio-economic status means 1 standard deviation above average

The random slopes for socio-economic status indicate differences between secondary schools in the relationship between socio-economic status and the score on the educational ladder. A strong negative association was found between the random slopes of socio-economic status and the random intercepts (r= -.68). This strong negative correlation indicates that more effective schools for the average student tend to have a more flat slope for socio-economic status. More ineffective schools tend to a steeper slope for socio-economic status. Differences between low and high socio-economic students in scores on the educational ladder are larger in these ineffective schools.

Model-based predicted probabilities for three typical schools are presented in Table 5.7 to illustrate the random slopes for socio-economic status. For each school the predicted probabilities of students with a low, average and high socio-economic status were given for possible scores on the educational ladder. The first of the typical schools is characterized by a steep slope for socio-economic status (b= -0.41) and a low effectiveness for the average student (b= 0.30). The second typical school can be characterized by an average slope for socio-economic status (b= -0.31) and an average effectiveness for the average student (b= 0.00). The third typical school is characterized by a relative even slope (b= -0.21) and high effectiveness for the average student (b= -0.30). It is apparent from Table 5.6 that the predicted cumulative probabilities for the more affluent students are fairly similar for these three typical schools. Larger differences arise between the typical schools for the student from less affluent families. For example, the chance of reaching a score of eight or higher on the educational ladder is .580 in the first typical school, .675 in the second typical school and .759 in the third typical school. There is a difference .179 (.759-.580) between the first and third typical school in the chance of students from less affluent families of reaching a score of eight or higher. In other words, students from less affluent families have a 17.9% larger probability of reaching a score eight or higher (diploma in pre-vocational education theoretical track or higher) in a school characterized by a relatively high intercept and a flat slope for socio-economic status. Furthermore, the differences between the more and less affluent students is the smallest in schools characterized by high intercepts and flat slopes.

5.3.3 Differences in value added between schools that provide education in different school tracks

Many schools for secondary education in the Netherlands provide education in multiple school tracks, with schools varying from providing education in one school

track to schools providing education in all school tracks. The educational ladder was designed as an outcome usable for all school tracks and provides the opportunity to estimate value added for a complete school. Therefore it is important to assess whether or not there is a relationship between the structure of the school (as measured by the school tracks in which education is provided) and the estimated value added. The estimated value added of secondary schools as well as the slope differences were drawn from Model 3 for further analysis.

No significant results were found between schools with different school composition for differences in intercepts, $F(4,62)=0.48$; $p=.75$. This implies that the composition of the school in terms of tracks cannot significantly predict differences in educational careers between schools for the average student. However, significant differences were found between schools with different school composition and their slope for socio-economic status, $F(4,62)=3.14$; $p=.02$. A similar picture arises for differences in slopes for prior achievement, $F(4,62)=10.86$; $p<.001$. The phenomenon of differential effectiveness for socio-economic status and prior achievement is clearly linked to differences between schools in the tracks in which they provide education. The schools that provide only education in the pre-vocational tracks have relatively flat slopes for both prior achievement and socio-economic status. These schools perform relatively well for the low prior achieving students and students from less affluent families. School with only pre-university education or all general tracks tend to have more steep slopes for both prior achievement and socio-economic status. Finally, schools that provide education in all tracks show average slopes for both prior achievement and socio-economic status.

5.4 Conclusion and discussion

This study set out to investigate the possibilities of estimating value added based on the educational careers of students. A number of advantages of a value added indicator based on educational careers of students can be formulated: (a) The societal significance of educational position as output measure, (b) the fact that a single indicator can be estimated for an entire school in a differentiated educational system, where not all schools provide education in all tracks. And (c) the expectation that value added based on educational careers leads to other incentives for schools than value added based on test scores.

For the first research question we assessed the school differences in effectiveness based on value added on the educational ladder. Modeling value added based on the Educational Ladder on the VOCL'99 cohort revealed small but significant differences in effectiveness between schools. The relative amount of between school variance in this current study appeared smaller than in previous studies in Dutch secondary education where models based on test scores were used (Thomas, 2001). However, given a 95% confidence interval surrounding each estimated school effect over 40% of the schools in this sample could be identified as significantly over- or underperforming. This indicates a value added indicator based on educational careers can discriminate between over- and underperforming schools, which can be considered as a precondition for an indicator in educational accountability.

To answer the second research question, differential school effects were assessed. For both prior achievement and socio-economic status differential school effects were found. In previous research differential school effects based on test scores have been found for prior achievement and socio-economic status in both primary and secondary education in the Netherlands as well as in other countries (Nuttall et al., 1989; Sammons et al., 1993; Thomas et al., 1997a; Veenstra, 1999; Gray et al., 2004). These differential school effects indicate that some schools are more and some schools are or less effective for some subgroups of students. For an accurate identification of over- or underperforming schools in the context of educational accountability it seems inadequate to simply estimate one value added score for the average student, as some groups of students based on prior achievement or socio-economic status benefit more from attending particular schools. This implies that in educational accountability, value added should be estimated for several subgroups for prior achievement and socio-economic status to get a more detailed and adequate impression of the efficiency of a school.

With respect to the third research question, the phenomenon of differential school effects for socio-economic status and prior achievement are associated with the composition of the school in term of tracks in which the school provides education. The schools that provide only education in the pre-vocational tracks have relatively flat slopes for both prior achievement and socio-economic status. This means that these types of schools perform relatively well for the student from less affluent families and students with low prior achievement scores. Schools providing education in the higher school tracks, such as general secondary education and pre-university education are characterized by steeper slopes for prior achievement and socio economic. This implies that these schools perform relatively well for the more affluent and high prior achievement students. Average slopes were found for schools that

provide education in all school tracks. In general, the specialization of schools with respect to their student population is reflected in their performance as measured by value added on educational careers of students. With respect to the usefulness of value added based on educational careers of students it is questionable whether schools with different school compositions can be compared due to the association between differential school effects and school composition. This implies that, at best, the value added based on the educational ladder can be used to compare schools in groups with the same structure with respect to school tracks.

Although including value added on the educational ladder besides indicators based on test scores might improve robustness against strategic behavior, the ordinal character of the educational ladder and the associated methods of statistical analysis result in scores that are difficult to understand for a non-statistical audience. Transparency is one of the criteria for the usefulness of indicators in educational accountability. Therefore, the complex ordinal modeling of value added based on the educational ladder raises further questions with respect to the usefulness as (a) How are the data to be presented? (b) What guidance is needed to help the target audience? (Reflection of Rosemary Butler, Department of Health UK, in Goldstein and Spiegelhalter, 1996).

A number of limitations of this study need to be considered. Attrition analysis revealed some bias on the student level, which might lead to some bias in the estimated value added of individual schools. This would be problematic if such an indicator would be implemented in educational accountability systems. However, the current study has a strong explorative character in which the aim was to explore possibilities and validity of modelling value added on educational careers.

Furthermore, the school where students participated in fourth grade or the last known school for the students that dropped-out was used in this study. However, in the four years of secondary education, covered in the current study, students might have attended multiple schools due to student mobility. One of the reasons of student mobility in secondary education is intermediate upward or downward mobility between schools, since not all schools provide education in all school tracks. Previous studies in the effects of modelling mobility of students on school effects based on test scores revealed that assigning students to the last school will result in an underestimation of the between school variance (Goldstein et al., 2007) and some bias in the estimated school effects (Leckie, 2009). In the light of an accurate estimation of value added of secondary schools it might be worthwhile to investigate the effects of

modelling student mobility on the estimated value added based on educational careers of students.

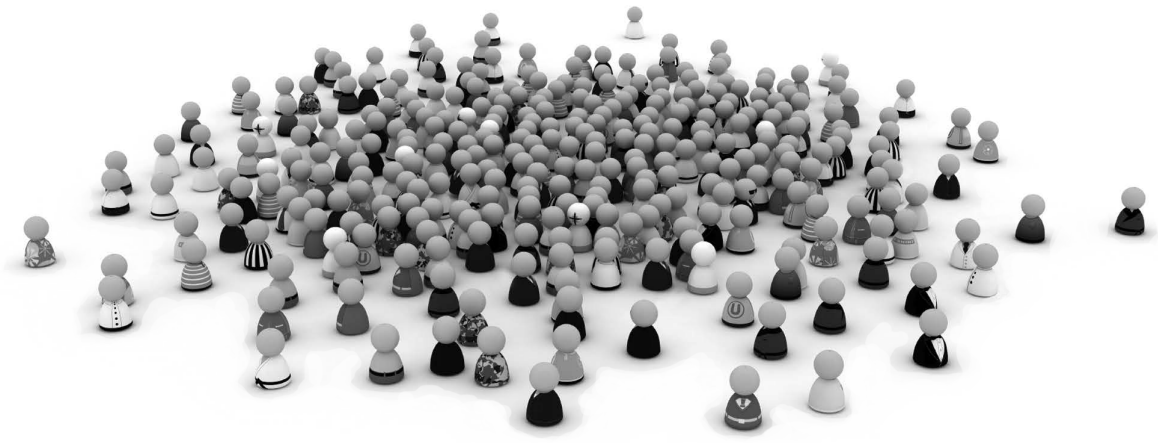
Ideally, one would like to compare the current value added model based on the educational ladder with existing performance indicators based on test scores. Due to differences in the content and level of the examinations between tracks in differentiated educational systems, such as the Dutch system, value added based on examination scores can only be estimated for tracks within schools. Differences of the level of inference (schools or track within schools) of these indicators make this comparison problematic.

This current study shows a number of supposed advantages of value added based on educational careers of students, although the differential school effects and complex statistical modelling remain important limitations of this indicator. Value added based on test scores and value added based on educational careers provide opposite incentives for schools if they are used in educational accountability. Value added based on test scores provide schools with the incentive to place students in lower tracks where they reach high test or examination scores, while value added based on educational careers provide schools with the incentive to place student in higher tracks where they might get lower test or examination scores. Strategic behaviour to artificially enhance one of the indicators will be at the expense of the other indicator. Therefore, it might be considered to use value added based on educational careers and test or examination scores as complementary indicators in educational accountability, because both indicators provide valuable information concerning the effectiveness of schools and tracks despite several flaws. Using multiple indicators might be beneficial for the robustness of an accountability system with respect to strategic behaviour (Koretz, 2003).

Chapter 6

Value Added as an Indicator of Educational Effectiveness in Dutch Senior Secondary Vocational Education

This chapter is based on:
Timmermans, A. C., Bosker, R. J., Doolaard, S., & De Wolf, I. (2012, In press). *Value added as an indicator of educational effectiveness in Dutch senior secondary vocational education.* Journal of Vocational Education and Training, DOI: 10.1080/13636820.2012.727849



Abstract

This study investigates the possibilities of estimating value added as a performance indicator in senior secondary vocational education. Value added is interesting in this context because it is considered as a reliable tool for comparing the effectiveness of educational institutions. Although value added indicators have been developed since the 1980s for both primary and secondary education, the research on school effectiveness has largely neglected vocational education because of its complexity. For estimating value added in this study, data concerning almost 90,000 students in Dutch senior secondary vocational education are used. Factors such as ethnicity, living in problematic neighbourhoods, and students' prior educational attainment appear to be significant predictors of student outcomes. The results indicate considerable differences in the effectiveness among clusters of training programmes, whereas there are hardly any between the educational institutions. Of the total variance among the student outcomes, 14% is related to the training programme clusters.

6.1 Introduction

Comparing the performances of educational institutions based on student outputs has grown in popularity. In several countries and states performance indicators are used in educational accountability systems and league tables. It is therefore crucial that these indicators are as reliable and valid as possible.

Unadjusted averages of student performance, for example average grades or pass rates, are insufficient as indicators of the performance of educational institutions because they also include factors of learning which go beyond the control of the school (Meyer, 1997; Webster et al., 1998). In other words, the unadjusted averages are partly determined by the characteristics of the student population, while the influence of these student characteristics is unevenly distributed among the educational institutions (Hill et al., 1996).

A more valid estimation of school performance can be achieved by taking the institutional differences regarding the intake of students into account. This approach is usually called value added. Prior educational attainment, background characteristics of the students, and school composition often serve as control variables in the estimation of value added. When estimating value added, usually a multilevel regression model is set up where students are nested within institutions, using control variables to predict the students' final achievement. The average difference between the expected and the actual performance at the level of the educational institutions is then used as the estimate of the value added.

In the tradition of school effectiveness research, value added indicators have been developed since the 1980's to assess the differences in the effectiveness of educational institutions (Raudenbush et al., 1986; Aitkin et al., 1986; Willms et al., 1989; Hill et al., 1996; Meyer, 1997; Goldstein, 1997; Bosker et al., 2001). Most of this research has focused on primary and secondary education. Our study aims at value added as a quality indicator in senior secondary vocational education, an educational sector of great interest to both the general public and policy-makers (Van den Berghe, 1996; Van den Berghe, 1997; Coates, 2009a; Coates, 2009b). Studies on the development and methodology of value added indicators in the sector of senior secondary vocational education (Harmon, 1992; Armstrong & McVicar, 2000) or higher education (Yunker, 2005; Rodgers, 2007) are scarce. The great variety of its

student population and its complex structure including multiple training programmes for different degrees, make senior secondary vocational education a challenging sector in terms of developing quality indicators.

6.1.1 Research questions

This study deals with the educational effectiveness of an educational sector that is under-represented in school effectiveness research. The aim of our study has been to develop value added estimates for institutions and training programmes in senior secondary vocational education, with a specific focus on the usability as performance indicator for educational accountability. In this paper we explore a possible methodology for estimating the value added of educational institutions in senior secondary vocational education, focusing on the relative differences among educational institutions with respect to the success of their students. We have formulated the following research questions:

1. Should accountability systems, based on performance indicators such as value added, focus on the level of training programmes or on the level of educational institutions in senior secondary vocational education?
2. Can underperforming and over-performing educational institutions and training programmes be identified using a value added model as a performance indicator?
3. Which control variables are relevant in estimating the value added of senior secondary vocational education?
4. To what extent are there differences in the identification of over- and underperforming training programmes between a value added indicator and a model which does not control for student characteristics?

The first section of this paper describes the methodologies proposed for estimating value added in vocational education in the literature, followed by a short overview of the Dutch educational system for senior secondary vocational education in particular. Next, we describe the design of our study. Finally, we present and discuss the results of our empirical analysis.

6.1.2 Estimating value added in vocational education

Generally, the value added of schools in primary and secondary education is estimated based on achievement or examination scores. In senior secondary vocational education, however, measuring the effectiveness of the educational institutions is far more complicated. Final achievement based on examination grades is not an appropriate measure in this educational sector because no national examinations are available. Instead, outcomes such as diplomas, credits, or pass rates have to be used. Furthermore, the great variety of training programmes for different degrees implies a more complex hierarchical structure of students, programmes, sectors, and educational institutions.

In most previous research, single level statistical models have been used to estimate value added in vocational education, such as single level ordered probit regression models (Armstrong et al., 2000; Rodgers, 2007), single level multinomial logit regression models (Rodgers, 2005) and multiple regression analysis on the institutional level (Yunker, 2005). These studies did not explicitly model the hierarchical structure of the data which consists of students (lowest level) nested within training programmes or institutions (higher levels). An adequate multilevel analysis is considered to be an important feature of value added modelling (Aitkin et al., 1986; Hill et al., 1996; Goldstein, 1997). The previous studies did not provide an indication of the differences among educational institutions or training programmes in terms of their value added.

Compared to primary and secondary education, the previous studies in vocational education have shown very similar results with respect to important student characteristics. They indicate a considerable relationship between students' prior educational attainment and subsequent success in vocational education (Armstrong et al., 2000). Furthermore, it has been argued that for an accurate analysis of value added, exogenous factors, such as ethnic origin and socio-economic background, should also be taken into account (Rodgers, 2005; Rodgers, 2007).

6.1.3 Dutch Senior Secondary Vocational education

Senior secondary vocational education in the Netherlands provides education for almost half of the Dutch students in further education. A schematic overview of the Dutch educational system is presented in Appendix I. Senior secondary vocational education offers a choice of approximately 700 training programmes for 500,000 students from the age of 16. These training programmes can be followed on four

levels via two different routes (Ministry of Education, 2007). There is a full-time college-based trajectory and a part-time work-based alternative, which combines part-time education with an apprenticeship in a company. The training programmes differ in length varying from 1 to 4 years, while not all of them are available at each level or via each route. All institutions in Dutch senior secondary vocational education together provide in total approximately 11,000 training programmes.

Regional training centres (ROCs) offer senior secondary vocational education in three sectors: engineering & technology, economics, and health & social care. Education in the sector agriculture, natural environment, and food technology is provided by agricultural training centres (AOCs). In total, there are 68 educational institutions for senior vocational education in the Netherlands. These institutions have clustered their training programmes based on their content. The clusters can be regarded as further specifications of the four sectors. They contain training programmes on different levels and of different length. Examples are "Transport and Logistic", "Graphics and Media", "Economy and Administration", and "Building and Infrastructure Contractor". However, not all institutions provide training programmes from all clusters. The largest educational institution offers training programmes from 14 different clusters, whereas some agricultural and small specialized training centres in the technical sector only provide a single cluster of training programmes.

6.2 Method

6.2.1 Subjects

The empirical analysis in this study is based on national student level data in senior secondary vocational education. In total, approximately 200,000 students left the publicly funded secondary vocational institutions in the 2007/2008 college year. About 16% of these students moved to other institutions in senior secondary vocational education, while 20% continued their studies by opting for other educational trajectories, especially universities for applied sciences (or higher vocational education). The majority of the students in the sample (63%) left the publicly funded educational system either with or without a diploma.

The analyses of the student population were based on the following criteria. All students had to have graduated or dropped-out in the college year 2007-2008 after following the college-based route (Dutch abbreviation BOL). Furthermore, information for all the covariates to be used had to be available. The total sample that

met these criteria consisted of 87,980 students, distributed among 442 clusters of related training programmes provided by 68 educational institutions.

6.2.2 Variables

Final achievement was the dependent variable and was measured by the level of the diploma obtained. This variable was used in five ordinal categories: 0 = no diploma obtained (42.3% of the students), 1 = diploma on the level of assistant (2.9%), 2 = diploma on the level of basic vocational training (11.6%), 3 = diploma on the level of professional training (8.7%) and 4 = diploma on the level of middle management training (34.6%). The category "diploma for middle management training" was used as the reference group in the analysis.

We operationalized the prior attainment variable, referring to the students' previous educational attainment in other institutions for vocational, secondary or higher education, into six ordered categories. This distinction in six categories is based on the requirement for students to attend education at specific levels. The reference group is the group of students that is allowed to attend training programs on the assistant level (1). The other groups are based on requirements for basic vocational education (level 2), professional training (level 3), middle management training (level 4), and universities for applied sciences and universities

We used background variables at the individual level as covariates. These included socio-economic status, measured by living in problematic neighbourhoods, ethnicity, and special educational needs. About 20% of the students lived in a problematic neighbourhood, which is defined as an area with a relative large number of residents living from social services or with a very low income, and a relatively large number of foreign residents. Ethnicity is a categorical variable, including Dutch students (70.3%), western foreign students (6.2%) and non-western foreign students (23.5%). We used the group of Dutch students as the reference group in the analysis. The covariate 'special educational needs' was measured by verifying whether the students had received additional support during secondary education (11.3%). The criteria for receiving additional support (Dutch abbreviation: lwoo) were based on whether students were lagging behind, low to moderate intelligence, and/or possible social and emotional problems.

6.2.3 Methods of analysis

In the analysis, the dependent variable is the level of diploma obtained, while several student characteristics are used as predictors. The students' prior educational attainment, ethnicity, (problematic) neighbourhood, and special educational needs serve as predictor variables. The differences between the prediction by the model and the actual performance of the students on the level of the training programmes (level 2 residuals) and the institutions (level 3 residuals) give an indication of the value added by the training programmes respectively the institutions. For the analysis of the ordinal dependent variable, such as the obtained diploma level, a two-level (students in training programme clusters) and a three-level (students in training programme clusters in institutions) multinomial ordered logit model were estimated using the MCMC algorithms in MLwiN (Rasbash et al., 2009; Browne, 2009). This model is also known as the multilevel ordered logistic regression model or multilevel proportional odds model (Hedeker, 2008).

Parameters resulting from MCMC algorithms were preferred over quasi-likelihood methods because they yield less biased estimates in the multilevel logistic regression analysis, especially in the case of estimating the random effects variance (Browne et al., 2006). This is especially important for the analysis of value added where the focus lies on the variance in residuals on the level of training programmes and institutions. We used the Deviance Information Criterion (DIC), a combined measure of model fit and model complexity, to compare the model fit of the models estimated based on MCMC algorithms (Spiegelhalter et al., 2002). Models with smaller DIC values are to be preferred over models with larger DIC values. A difference of five points between the models is considered as a substantial improvement of model fit.

6.3 Results

6.3.1 Which level is appropriate for estimating the value added of senior secondary vocational education?

For the estimation of the value added indicators we used two multilevel models: one model with two levels and one model with three hierarchical levels of nesting. To investigate the differences in effectiveness between the training programme clusters and the institutions we estimated a three level model containing students (level 1) nested within training programme clusters (level 2) and training programme clusters

nested within educational institutions (level 3). The estimated effects or value added of the educational institutions is presented in Figure 6.1. In this plot, each triangle depicts the estimated value added of an educational institution. Connected to this triangle is the 95% confidence interval of the estimated effects. On the horizontal axis, the educational institutions are ranked from most to least effective. Each of the confidence intervals for the value added estimates of the educational institutions includes zero. This means that none of the institutes can be distinguished from average. There appears to be only a very small amount of variation in the effectiveness among the institutions, which means that the educational institutions for senior vocational education are quite similar in their effects on student outcomes. Further statistical tests revealed that the three level model, which includes educational institutions, does not necessarily have to be preferred over the more simple two level model (DIC = 214,397.74 for the three level model and DIC= 213,991.77 for the two level model).

Figure 6.1

Caterpillar plot of the estimated value added and the associated 95% confidence intervals on the level of educational institutions

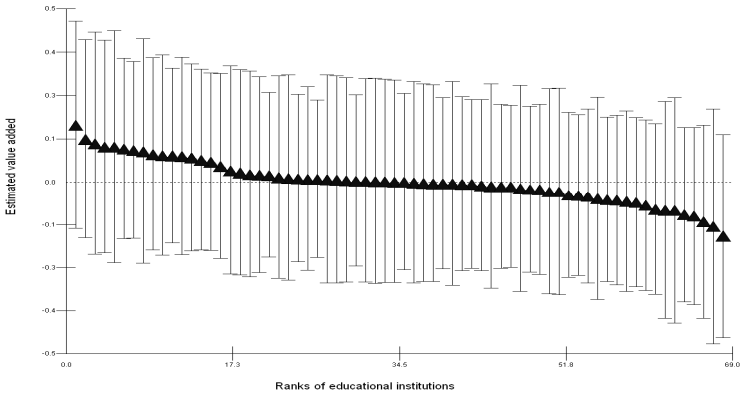


Table 6.1

Results from the MCMC estimation of the two level multinomial ordered logit gross effect model with diploma middle management level as reference category.

	Coefficient (mean)	Standard Error	Coverage Interval (95%)
Fixed part			
No diploma	-0.275***	0.037	[-0.349, -0.203]
Diploma assistant level (1)	-0.149***	0.037	[-0.224, -0.0.78]
Diploma basic vocational education level (2)	0.345***	0.037	[0.270, 0.416]
Diploma professional training level (3)	0.733***	0.038	[0.658, 0.804]
Random part			
Level of training program variance	0.654	0.055	[0.554, 0.768]
Model fit			
DIC	221799.994		
Units: Students	87980		
Units: Training programs	442		

*** $p<0.001$ (two-tailed). Coefficients are reported on the log odds scale.

Full model specifications of the two level model gross school effects model are given in Table 6.1. The model specifications of the value added model are given in Table 6.2. In this model about 14% of the variance appeared to be located at the level of the training programme clusters. In Dutch secondary education differences between schools are of a similar magnitude (Luyten, 1998; Veenstra, 1999).

Because the category "diploma on middle management training" (level 4) is the reference group in the model a positive coefficient in the table implies an increase in the probability of obtaining no diploma increases and a decrease in the probability of obtaining a diploma on the level of middle management. A negative coefficient in the model implies an increase in the probability of obtaining a diploma on the higher levels in vocational education. For interpretation purposes the figures and the table in the paper have been mirrored, with the effect that schools with a positive residual are assessed above average and schools with a negative residual below average.

Table 6.2

Results of the MCMC estimation of the two level multinomial ordered logit value added model with diploma middle management level as reference category.

	Coefficient (mean)	Standard Error	Coverage Interval (95%)
Fixed part			
No diploma	0.363***	0.051	[0.271, 0.455]
Diploma assistant level (1)	0.491***	0.051	[0.399, 0.583]
Diploma basic vocational education level (2)	1.026***	0.051	[0.933, 1.118]
Diploma professional training level (3)	1.465***	0.051	[1.371, 1.557]
Western foreign students	0.338***	0.027	[0.284, 0.392]
Non-western foreign students	0.443***	0.019	[0.406, 0.480]
APCG	0.183***	0.019	[0.146, 0.219]
LWOO	0.016	0.021	[-0.025, 0.058]
Prior educational attainment			
Requirements for basic vocational education	-0.164***	0.032	[-0.227, -0.102]
Requirements for professional training	-0.665***	0.032	[-0.727, -0.603]
Requirements for middle management training	-1.325***	0.031	[-1.387, -1.264]
Requirements for higher vocational education	-1.454***	0.047	[-1.547, -1.362]
Requirements for university	-0.950***	0.198	[-1.338, -0.562]
Random part			
Level of training program variance	0.544	0.048	[0.458, 0.645]
Model fit			
DIC	213991.744		
Units: Students	87980		
Units: Training programs	442		

*** $p<0.001$ (two-tailed). Coefficients are reported on the log odds scale.

6.3.2 Can overperforming and underperforming training programme clusters be identified?

A plot of the estimated value added of the training programme clusters is presented in Figure 6.2. In this figure each triangle is the estimated value added of a training programme cluster in an educational institution ranked on the basis of its estimated effectiveness. The height of the triangle on the y-axis represents the estimated combined effect of the educational institution and the training programme cluster. The range of estimated effects lies between 2.2 and -3.3 on the log odds scale. The clusters of training programs are more variable in their effectiveness than institutions. Furthermore, there is considerable variation in the size of the confidence intervals between training programs. The size of the confidence interval is amongst others based on the number of students within a cluster of training programs.

Three groups of training programme clusters can be identified in Figure 6.2, namely (a) average training programmes, (b) overperforming training programmes, and (c) underperforming training programmes. A value of zero means an average value added. The largest group, the average training programmes, is formed by training programme clusters for which the confidence interval around its estimated effect (residual) includes zero. This means that the effectiveness of these training programme clusters cannot be statistically distinguished from average. A small group of training programmes performs better than the average. The confidence intervals around the estimated effects of the training programmes in this group are above zero on the log odds scale. These programmes can be distinguished as above average in terms of effectiveness or as over-performing. Simultaneously, a small group of training programme clusters can be considered as less effective or underperforming. The confidence intervals for these programmes are below zero. This small group of clusters might be of particular interest for accountability purposes, which are usually aimed at identifying the less effective training programmes. The caterpillar plot indicates that both overperforming and underperforming clusters can be identified, although the confidence intervals of the majority of the training programmes (49.1%, 217 training programmes) include zero. In total 136 (30.5%) training programmes can be identified as over-performing and 89 (20.4%) as underperforming.

Figure 6.2
Caterpillar plot of the estimated value added and the associated 95% confidence intervals on the level of training programme clusters

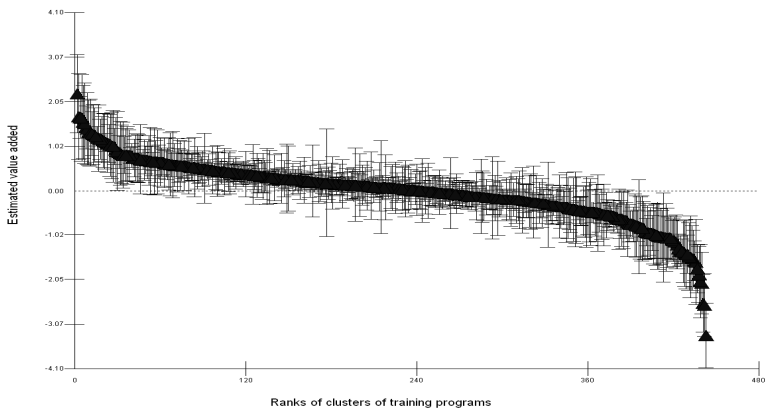
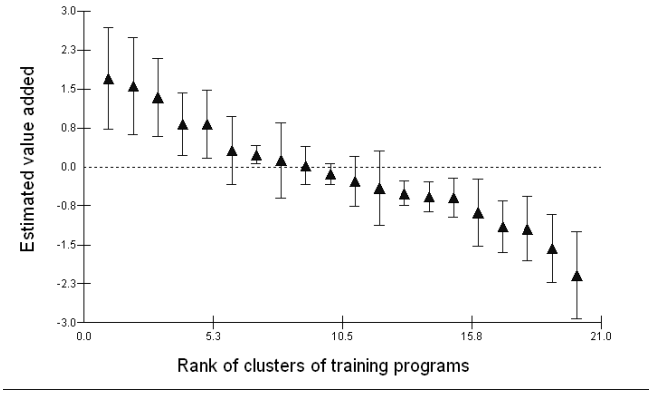


Figure 6.3
Caterpillar plot of the estimated value added and the associated 95% confidence intervals for the Graphics and Media training clusters in different educational institutions



6.3.3 Differences among training programme clusters regarding the predicted probabilities of obtaining diplomas

In Table 6.3, the model-based predicted probabilities are presented for training programs with different levels of effectiveness. These probabilities correspond to the triangles in the previous plots and give a more practical indication of the differences in effectiveness among clusters of training programmes. The predicted probabilities were calculated for both low- and high-performing training programmes with an estimated effect of one and two on the log odds scale. Furthermore, the number of schools between certain ranges on the log odds scale are given. There are remarkable differences in effectiveness among the training programmes, as can be seen from differences in predicted probabilities of obtaining diplomas. For example, in an underperforming cluster with a score of -1 on the log odds scale a Dutch student without special educational needs, living in a non-problematic neighbourhood, with a prior educational attainment required for a middle management training programme has a probability of .24 of obtaining a diploma on the middle management level (4). A similar student in a cluster with a score of +1 on the log odds scale has a probability of .70 of obtaining a diploma on the middle management level.

Table 6.3
Model-based predicted probabilities of obtaining diplomas for training programmes with different effectiveness levels[1]

Training programmes	Number of training programmes	No diploma	Diploma level 1	Diploma level 2	Diploma level 3	Diploma level 4
Very high (between 1.5 and 2.5[2])	6	.04	.01	.04	.04	.86
High (between 0.5 and 1.5)	84	.12	.01	.08	.08	.70
Average (between -0.5 and 0.5)	271	.28	.03	.13	.11	.46
Low (between -1.5 and -0.5)	68	.51	.03	.13	.09	.24
Very low (between -2,5 and -1,5)	13	.74	.02	.08	.05	.10

[1] For Dutch students without special educational needs living in a non-problematic neighbourhood, with a prior educational attainment on the level of middle management
[2] On the log odds scale

6.3.4 What are relevant control variables for estimating the value added of senior secondary vocational education?

In Table 6.4, information is presented from which the importance of the covariates for predicting success in senior secondary vocational education can be derived. Significant associations were found between students' success in vocational education on the one hand, and their prior educational attainment, neighbourhood, and ethnic origin on the other hand. No significant effects were found for students who had received extra educational support (lwoo) in the past. The relation between prior educational attainment and students' success is positive. Higher prior attainment leads to higher estimated probabilities of obtaining a diploma on a higher level. For example, a Dutch student with a prior educational attainment required for the basic vocational track has a probability of .21 of obtaining a diploma on the level of middle management training. A student with prior educational attainment required for professional or higher education has a predicted probability of .50 of obtaining a diploma on the middle management level.

From Table 6.2 it becomes apparent that both western and non-western foreign students are significantly more likely to leave the educational system without a diploma or with a lower diploma than Dutch students. The probability of leaving the educational system without a diploma is 9% (.37 - .28) higher for non-western foreign

students and 7% (.35 - .28) higher for western foreign students in comparison to Dutch students with similar levels of prior educational attainment. Similarly, the probability of obtaining diplomas for the higher levels in senior secondary education is lower for students living in problem neighbourhoods. These students have a 4% (.32 - .28) higher probability rate of leaving the educational system without a diploma compared to students living in non-problematic neighbourhoods with similar levels of prior attainment and the same ethnic origins. The probabilities of obtaining diplomas on levels 1, 2 and 3 vary only slightly among the different levels of covariates compared to the probability of leaving the system without a diploma or a diploma on level 4. This strongly depends on the levels of the other covariates.

Table 6.4

Model-based predicted probabilities of obtaining diplomas on different levels of covariates

	No diploma	Diploma level 1	Diploma level 2	Diploma level 3	Diploma level 4
Prior educational attainment[1]					
Requirements for assistant level	.59	.03	.12	.08	.19
Requirements for basic vocational education	.55	.03	.12	.08	.21
Requirements for professional training	.42	.03	.13	.10	.31
Requirements for middle management training	.28	.03	.12	.11	.46
Requirements for higher vocational education	.25	.02	.12	.12	.50
Requirements for university	.36	.03	.13	.11	.37
Ethnicity2					
Dutch students	.28	.03	.12	.11	.46
Non-western foreign students	.37	.03	.13	.11	.36
Western foreign students	.35	.03	.13	.11	.38
Problematic neighbourhood3					
Not living in a problematic neighbourhood	.28	.03	.12	.11	.46
Living in problematic neighbourhood	.32	.03	.13	.11	.42

[1] For Dutch students without special educational needs and living in a non-problem neighbourhood

[2] For students without special educational needs, living in a non-problematic neighbourhood, with a prior educational attainment on the middle management level

[3] For Dutch students without special educational needs, with a prior educational attainment on the middle management level

6.3.5 Differences in the identification of over- and underperforming training programmes between the value added and the unadjusted scores

To show the importance of controlling for student characteristics, a comparison was made between an empty model without any covariates and the value added model. An empty model can be seen as the unadjusted score or the gross effect of a cluster of training programmes. The full model specifications of the gross effect are presented earlier in Table 6.2. In Table 6.5, the clusters of training programmes are classified into the three groups depicted in the caterpillar plots, namely overperforming, average, and underperforming training programme clusters. Table 6.5 shows the classifications of both models into these three groups. The clusters of training programmes on the diagonal of the table are classified into the same groups by both models. In total, both models classified almost 80% of the clusters of training programmes into the same groups, which corresponds with an agreement of Kappa = 0.689; $N = 442$; $p < 0.001$. This result can be interpreted as a substantial agreement between the two models.

Table 6.5
Classifications of the training programme clusters into over-performing, average, and underperforming

		Value added model		
		Underperforming	Average	Over-performing
Unadjusted model	Underperforming	84	37	1
	Average	5	153	19
	Over-performing	0	27	116

Deviations from the diagonal show that the models also differ in their classification of some clusters. For example, according to the unadjusted probabilities of obtaining diplomas, 112 clusters are classified as below average. However, 37 clusters of this group can be identified as average after controlling for students' prior educational attainment and background characteristics, while one cluster even performs above average. For 20% of the training programme clusters a value added indicator would lead to a different classification. Therefore, the use of these indicators for educational accountability purposes could have important consequences for individual clusters of training programmes.

6.4 Conclusion and discussion

This study has explored the possibilities of estimating value added as an indicator of the performance of educational institutions in senior secondary vocational education, with a special focus on the context of accountability. We specifically addressed the educational effectiveness of senior secondary vocational education because this area is underrepresented in the research on school effectiveness. This paper should be seen as a first step in the development of a value added analysis in this educational field.

One of the significant findings of this study is that only a very small fraction of the variance in student outcomes is associated with the level of educational institutions. This does, however, not necessarily imply that educational institutions have no effect on the success of students in senior secondary vocational education, but that institutions are rather similar in their effects on students. It appears that the differences in the effects of the training programme clusters are neutralized at the level of the educational institutions. An institution can be effective for training programme A and less effective for training programme B, which brings the average effectiveness of the educational institution as a whole closer to zero. A significant amount of variance, however, has been found on the hierarchical level of clusters of training programmes. This is why these clusters can be considered as a more appropriate level to make inferences when using value added indicators in accountability systems. From an accountability perspective the advantage of the two level model is that it produces value added estimates based on students nested within training programmes, which makes it possible to compare the effectiveness of training programmes independently of institutions. For example, in this model the cluster of Transport and Logistics of institution A can be compared to the Agriculture training programmes of institutions B and C or to other training programmes in institutions A, B and C.

Value added indicators seem to grasp a part of the quality or effectiveness of training programs and offer the possibility of comparing training programs over educational institutions. About 30% of the training programmes studied can be identified as overperforming and 20% as underperforming. Furthermore, the predicted probabilities of obtaining a diploma differ considerably between underperforming and over-performing training programmes. Compared to the unadjusted scores 20% of the clusters are given different classifications in terms of underperforming, average, and overperforming, whenever a value added model is used. These differences could have important implications for individual clusters of training programmes, depending on which model – the unadjusted or valued added – is chosen in an educational accountability system. In addition, similar to previous

studies conducted in higher and vocational education (Rodgers, 2005; Rodgers, 2007) our research has shown that prior attainment, ethnic origin and socio-economic background are significant predictors of student success, that should be taken into account when estimating value added in senior vocational education Finally, whether or not students received any additional educational support during their secondary education appeared not to be a significant predictor of student success.

However, when comparing institutions or training programmes in terms of their effectiveness, the large confidence intervals associated with their estimated effects remain important and should not be disregarded (Goldstein et al., 1996; Leckie et al., 2009). Furthermore, this kind of indicators should be used in a balanced system of quality indicators and can be a valuable addition to sets of quality indicators as described by Van den Berghe (1997). An example of another indicator might be the time spend in an educational institutions as a measure of the efficiency of an educational institution.

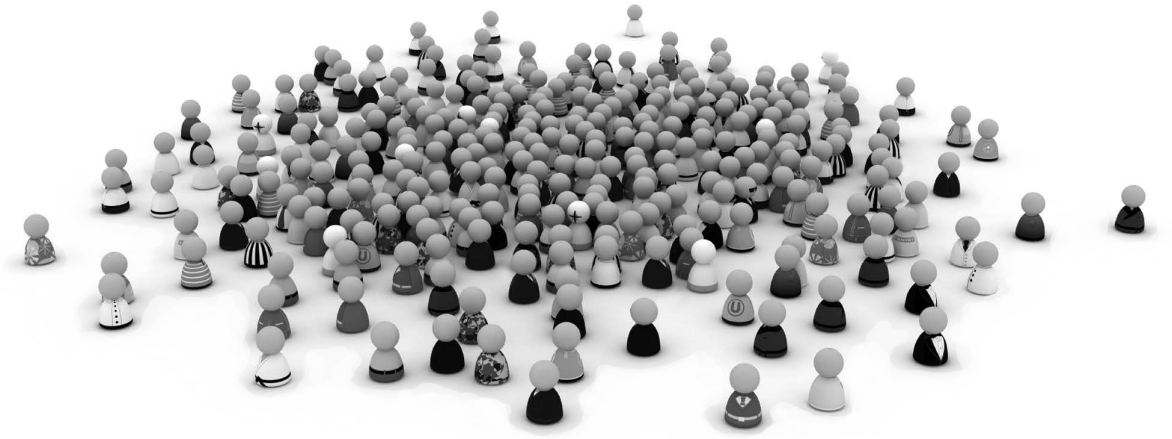
Nevertheless, these results should be handled with caution. Organizational and policy differences among educational institutions as well as among training programme clusters might, for example, lead to differences in the numbers of students obtaining diplomas on the several levels. Especially drop-out prevention policies might influence these results quite strongly. In some clusters of training programs, the education in different levels of similar training programs runs parallel. In case of drop out, students in these training programs may receive a diploma on a lower level, whereas in other training programs the students would leave and obtain no diploma. Finally, there is the issue of data limitations, such as missing values for prior attainment and newly developed training programmes, which might lead to an underestimation of the predicted probability of obtaining a diploma on the higher levels.

This study has produced many questions which require further investigation. Future studies should assess the stability of the estimated value added of educational institutions over time. The stability of value added estimates over time can be interpreted as a measure of reliability, since it is not expected that training programmes will have large fluctuations in their effectiveness over successive years (Willms et al., 1989; Luyten, 1994; Thomas et al., 1997b). In addition, it might be worthwhile to investigate the consistency of value added with respect to other possible outcome variables, such as credits earned or exam grades. Finally, before value added indicators can be properly used in accountability systems, their validity should be thoroughly investigated. Multiple methodologies have been used to

investigate their validity in other educational sectors, for example by making comparisons with other indicators of school quality (Yunker, 2005; Gorard, 2006; Van de Grift, 2009) or assessing more complex hierarchical structures to identify other contributing factors, such as the long-term effects of secondary schools, neighbourhoods, and student mobility (Goldstein et al., 1997; Goldstein et al., 2007; Leckie, 2009).

This article explored the possibility of estimating value added in senior secondary education. This might be of particular interest for any governmental body with accountability tasks, because more valid and fair estimations of the performance of educational institutions can be achieved by taking the institutional differences regarding the intake of students into account.

Chapter 7
Risk-based Educational Accountability in Dutch Primary Education



Abstract

A recent development in educational accountability is a risk-based approach, in which the intensity and/or frequency of school inspections vary across schools dependent on the risk level of a specific school. A risk-based inspection system is considered to be more effective because it enables inspectorates to focus on organizations at risk. In this article we assess which characteristics of primary schools are relevant in predicting which schools are "at risk" and how robust a risk model is over multiple cohorts based on an empirical analysis of 500 Dutch primary schools. At risk schools were defined as schools performing significantly below average on achievement and value added indicators. The composition of the school, previous performance, a systematic approach and evaluation of effects of extra care and monitoring the performance of students appear as the best predictors of underperformance of the primary schools in the sample. The results indicate that, if risk models for predicting underperformance of primary schools would be applied in the context of educational accountability, a large number of schools need further investigation to find nearly all underperforming schools.

7.1 Introduction

7.1.1 Risk-based school inspections

In most European countries, Inspectorates of Education assess the quality of public schools. In general, the aims of inspectorates are to guarantee a minimum quality level and to improve the quality of schools. The main instruments of inspectorates are school inspections, in which inspectors asses the quality of (the educational processes in) schools. Recently, some of the European inspectorates introduced a risk-based strategy to improve the efficiency and effectiveness of school inspections (Ofsted, 2011; De Wolf & Verkroost, 2011). A risk-based inspection system is considered to be more effective because it enables inspectorates to focus on organizations at risk (Sparrow, 2000). These are organizations with a high risk of noncompliance or underperformance. Inspections at these non-committing or underperforming organizations are more effective than inspections at well performing organizations. Risk-based inspections are also considered to be more efficient than traditional inspection systems. The efficiency gain lies in a less intensive inspection regime for well performing organizations.

The American labor inspection (OSHA) was the first inspection organization that developed a combination of risk(management) and inspections (Sparrow, 2000). The OSHA stressed that the increasing number of inspections did not seem to have any effect on reducing risks, which they ascribed to the fact that there are simply not enough inspectors to control all the American enterprises at a certain level. OSHA proposes a new strategy: Risk-based Inspection (RBI). The idea behind this is as simple as attractive; focus on the high-risk enterprises (instead of all enterprises). Sparrow's book inspired many inspection and audit organizations to introduce risk analyses in order to focus the inspection effort.

In educational accountability, a risk-based strategy implies that the intensity and/or frequency of school inspections vary across schools dependent on the risk level of a specific school. Schools at risk are inspected more frequently and more intensively than schools with a low chance of being a failing school. A risk analysis enables inspectorates to do in-depth school inspections at schools at risk. When risk levels are monitored at a continuous or frequent basis, risk models can also function

as an 'early warning system'. An increase in risk level of a specific school can be used to decide for an extra inspection.

7.1.2 Risk analysis

Risk analysis started in the beginning of the previous century as a technique to predict life expectancy (including human health) for insurance and banking institutions. It has been developed in a wide tradition, especially in epidemiology, econometrics and technical mathematics. Modern risk analysis has its roots in probability theory and the development of scientific methods for identifying the causal link between adverse health effects and possible hazards (Molak, 1997). Risk analysis -or risk assessment- is nothing more than a systematic analysis of risks. Or, as the Australian government defines it: 'risk analysis is a systematic use of available information to determine how often specified events may occur and the magnitude of their consequences' (Standards Association of Australia, 1999). Numerous of these kinds of risk analyses exist, most of them applied to safety risks and risks in project management (Keeney & von Winterfeldt, 2011; de Jong, 2012).

The definition of a risk or an adverse effect is a crucial first step in risk analysis (Standards Association of Australia, 1999). In risk analysis, defining an adverse effect is considered to be a value judgment (Molak, 1997). Examples of well-known adverse effects in the tradition of risk analysis are death or diseases, failure of nuclear power plants and loss of investments. Within the context of educational accountability, we define two adverse effects for schools. In the first place low performance of students within schools at the end of a formal stage of education can be considered as an adverse effect, as students need to reach certain standards for successful careers in society or in further education. The second adverse effect for schools is a low growth in performance of their students during this formal stage of education. In this paper, we therefore define a school at risk as a school with low academic achievement and/or a school with low growth in achievement. This latter is measured by value added. Value added has been adopted as an indicator of the performance of schools in many educational accountability systems (Sanders et al., 1994; Sanders, 2003; Ray, 2006; Betebenner, 2007; Betebenner, 2009; Ofsted, 2010). Value added has become a concept for a collection of statistical models in which the growth of students in schools is assessed while controlling for differences in intake of students between schools.

A risk analysis is the assessment of characteristics or (risk) factors that relate to an adverse effect. Furthermore, it tries to estimate at which levels of these factors the adverse effects become more prevalent. An example of this latter is the critical amount of exposure to hazardous chemicals at which the majority of the people tend to get sick. Finally, risk analysis tries to answer to which extent it is possible to accurately predict the prevalence of risks or adverse effects.

Two methods of risk assessment are common: an empirical method based on data and methods based on expert judgments (Molak, 1997). The empirical models assume that, based on historical data, one can establish the probability of adverse effects (Molak, 1997). Regression models on retrospective cohort data are used frequently in risk analysis, to predict a risk level at time point *t* based on data at time point and *t-1* and before. It is assumed that the factors that predict risks or adverse effects at time point t also predict possible risks in the future (*t+1*). Epidemiological studies in which persons with a certain disease and persons without that disease are compared in exposure and other factors are examples of these empirical models of risk analysis. In this current paper, we apply an empirical method to assess which characteristics of schools are associated with underperformance of schools.

Risk analysis requires good, reliable and integral data. This condition is so important that many attempts of introducing risk management and applying risk analyses just fail due to lack of data. Or, as Hulett and Preston (2000) put it: "collecting better data is the best way to improve your risk assessment" (Hulett & Preston, 2000). The transition to risk-based school inspection systems became possible by an increase in available and standardized data at the school level in most European countries. These datasets make it possible to calculate risk levels for individual schools. In most countries, available school level data are data on test scores, past performance, student and teacher characteristics, signals concerning children's safety within schools and practices and policies within schools.

The consistency of the results of risk analyses over multiple studies is considered to be an important requirement before risk models can be applied in practice (Gibb, 1997). Previous studies within the tradition of educational effectiveness research have merely focused on identifying characteristics of effective schools instead of underperforming or failing schools (Reynolds & Teddlie, 1999b). Whilst sets of characteristics of schools that appear very effective over multiple studies have been formulated, it lacks adequate descriptions of underperforming schools. It is the group of underperforming schools and their characteristics that are of special interest in educational accountability and risk analysis.

7.1.3 Research questions

The aim of the current study is to investigate the possibilities of risk analysis in education in the context of educational accountability. The present study answers the following research questions:

1. Which factors and which levels of these factors are relevant in predicting "at risk" schools in primary education?
2. How robust is a risk model over multiple cohorts?

Answers to these questions provide insight into the stability of measures of value added by schools over time and the extent to which underperformance can be predicted. Together, this provides an indication of the usefulness of a measure of added value in risk assessment. The second section of this study gives an overview of the data and methods used in this study. In the third section, the empirical results will be presented. Implications of the results for both educational effectiveness research and educational accountability will be discussed in the last section of this paper.

7.2 Method

7.2.1 Subjects

The first dataset used for the analysis in this empirical study is a student level dataset derived from the Monitoring and Evaluation system of CITO, the Netherlands Institute for Educational Measurement. This system offers schools and teachers the possibility to monitor the progress of their students during primary education, through a set of tests for several subjects, a registration system and tools for identifying specific learning problems and remediation guidance. The data for the current study contained test scores for reading comprehension and other subjects of students in Dutch primary education. Furthermore, the student's age and gender is recorded in the data. Other student background characteristics, such as the socio-economic status, are not available in the dataset. This study focusses on reading comprehension, as this subject is considered as an important prerequisite for further learning, because most information comes in the form of written texts. Students need to be able to comprehend what they are reading for good performance in reading tasks or tasks that require reading, as most tasks in education. Reaching certain levels

of reading comprehension is important for an adequate future in secondary and further education.

For reading comprehension three combined cohorts of students from grade three until grade five are constructed. To ensure a reasonable reliable estimation of the value added of a school, the students from cohort 2003 and cohort 2004 are combined in cohort 2003/2004, which leads to a larger number of students per school and a larger reliability of the schools' estimated value added. The other cohorts are constructed in similar ways. The number of students and schools in each of these cohorts are presented in Table 7.1 for reading comprehension. These students and schools were selected based on the following criteria: 1) identification variables had to be available at the student level and the primary school level, 2) test scores had to be available for the student from grade three until grade five on the reading comprehension tests, and 3) test scores had to be available for schools for both individual cohorts. Covering a longer time span for the estimation of value added, through including grade two or grade six, results in a great loss of both students and schools.

Table 7.1

Samples for reading comprehension

Cohort		2003/ 2004	2004/ 2005	2005/ 2006	2006/ 2007	2007/ 2008	2008/ 2009	2009/ 2010	2010/ 2011
2003	Students	7,693	7,316	7,020					
	Schools	265	264	257					
2004	Students		8,458	7,904	7,514				
	Schools		299	290	288				
Cohort 2003/2004: 15,195 students and 262 schools									
2005	Students			8,881	8,256	7,933			
	Schools			314	304	304			
2006	Students				10,061	9,383	9,072		
	Schools				338	334	330		
Cohort 2005/2006: 17,886 students and 314 schools									
2007	Students					11,390	10,645	8,946	
	Schools					375	370	314	
2008	Students						12,755	10,496	6,065
	Schools						416	356	220
Cohort 2007/2008: 22,815 students and 371 schools									

7.2.2 Instruments and variables

Of focal interest in the current study is the performance of schools in the domain of reading comprehension. For reading comprehension tests are used from the CITO Monitoring and Evaluation system. These tests are administered by the schools to monitor the performance of their students. Descriptive statistics of the variables used in the analysis are presented in Table 7.2.

Reading comprehension. In grade three, four and five, the level of reading comprehension of students is measured by taking a grade specific test developed by CITO, the Netherlands Institute for Educational Measurement. For each test there is a paper and digital version. In these tests, the comprehension, interpretation and reflection of written texts is measured by 50 items. The tests contain different types of texts (e.g. informative, fiction) and different types of genres (e.g. poems, letter, story, article). The questions within the tests can be related to the comprehension of the content of the text and the structure of the test. The scores of the students on these reading comprehension tests can be converted to a single latent one-dimensional reading comprehension scale (Feenstra, Kamphuis, Kleintjes, & Krom, 2010). The possibility to convert the test results to a single scale offers the possibility to monitor the progress that students or groups of students make over time. In the period between 2003 and 2010 a number of schools changed to newer versions of the reading comprehension tests in the Monitoring system. However, the scores on the new version are converted to the same latent scale. The reading comprehension tests in the Monitoring and Evaluation system have reliabilities between .84 and .93 for the paper versions and between .83 and .93 for the digital versions.

Time. The time variable for the growth models is operationalized for each cohort by calculating the proportion of time in years between the exact date on which the test was made and the end of grade five (1st of June).

Cohort. This variable indicates whether a student in the combined cohorts belonged to the first year (for example cohort 2003 from the 2003/2004 cohort) or to the last year.

Furthermore, a school level dataset with characteristics of the schools in 2009 is available from the Dutch Inspectorate of Education to investigate which factors are relevant in predicting "at risk" schools. This dataset contains variables concerning the population of teaching and supporting staff, school board, and student composition, size of the school, province, vision and religion and practices and processes within

schools. A number of possible predictors of underperformance of schools are derived from school inspections, standardized visits to assess the quality of primary schools. During these school inspection, inspectors visit lessons and interview teachers and principals. They use a standardized method and framework to assess the quality of schools. In this paper, we use the main inspection results, i.e. the assessment results of various aspects of educational quality. The most recent inspection results are used, as not all schools are inspected every year and with the complete framework. The following variables are used in the risk analysis:

- *All goals for Dutch language and Mathematics are covered in the curriculum* and *Schools with a high proportion of students from low educated parents provide adaptive teaching arrangements for Dutch language* are variables that indicate whether school provide an adequate curriculum to their students. Both variables are measured in two categories, namely sufficient (coded as 1) and insufficient (coded as 0).
- The variables *Teacher explains clearly*, *Students are engaged with the learning activities* and *Realization of a task-oriented working atmosphere* provide information on the practices and processes within the classroom. Usually, these indicators are measured through a number of classroom observations. Therefore, these indicators provide a general indication of the quality of classroom practices within schools and are not directly related to reading comprehension. Similar to the previous variables these indicators of classroom practices are measured in two categories, namely sufficient (coded as 1) and insufficient (coded as 0).
- Two variables are related to the provision of extra care for struggling students or students with learning disabilities, namely *The use of a systematic approach of providing extra care* and *Regular evaluation effects of extra care*. Similar to the previous indicators, these variables give a general indication of the quality of the provision of extra care. These indicators are measured in two categories, namely sufficient (coded as 1) and insufficient (coded as 0).
- The variables *The use of a coherent system of instruments and testing for monitoring progress of students*, *Teachers monitor the progress of students systematically*, *Teachers monitor the progress in the development of students systematically*, *Annual evaluation of the performance of students* and *Regular evaluation of the learning process* are all indicators of the evaluation and monitoring of schools. Similar to the previous variables these indicators of classroom practices are measured in two categories, namely sufficient (coded as 1) and insufficient (coded as 0).
- The most recent *inspection judgment* gives a general indication of the quality of the school. The inspection judgments are based on the performance of schools as measured by test scores during and at the end of primary education and an

assessment of the quality of the process of education during school inspections. This indicator is measured in three categories. The majority of the schools are considered sufficient and a small number of schools in this sample are considered inadequate or very poorly. Schools are judged as inadequate when the performance of their students on tests or the quality of the process is considered insufficient. Schools are judged as very poorly when the performance of their students on tests and the quality of the process is considered insufficient. Inadequate and very poor schools are visited more regularly.

Table 7.2a

Descriptive statistics of the main inspection result

Variables	***N***	**% of under-performing schools**
Sufficient performance of students at the end of grade 6 given the student population of a school	481	2.9
All goals for Dutch and Mathematics are covered in the curriculum	460	3.5
Schools with a high proportion of students from low educated parents provide adaptive teaching arrangements for Dutch language	460	4.8
Teacher explain clearly	462	3.7
Students are engaged with the learning activities	461	3.5
Realization of a task-oriented working atmosphere	461	2.6
The use of a coherent system of instruments and testing for monitoring progress of students	483	7.9
The use of a systematic approach of providing extra care	483	31.7
Teachers monitor the progress of students systematically	451	8.6
Teachers monitor the progress in the development of students systematically	181	33.1
Regular evaluation effects of extra care	470	39.8
Annual evaluation of the performance of students	478	32.2
Regular evaluation of the learning process	474	32.1
Inspection judgment	461	
Satisfactory		95.9
Inadequate		3.7
Very poorly		.4

- *Number of staff* and *Total number of students* are variables indicating the size of the schools. These variables are measured by the total number of employees and the total number of students within a school.
- *Staff until 30 years of age* and *Staff from 56 years of age* give an indication of the experience of the staff within schools. These variables are measured by the percentage of staff in these age groups.
- *Female staff* is measure by the percentage of female staff within a school.
- *Growth of staff* and *Growth of students* give indications of possible growth of the school. This variable is measured in the percentage of growth of the number employees within the last year.
- The variables *Staff intake from outside primary education, Staff intake from other primary schools, Staff leaving outside primary education* and *Staff leaving to other primary schools* give more detailed information of staff movements with respect to the school. Especially, staff intake from outside primary education might indicate new more inexperienced teachers or other staff. These variables are measured in percentages within the last year.
- The variables *Percentage of supporting staff, Percentage of part timers* and *Percentage of management* give some indications of the composition of the staff within a school with respect to different positions. These variables are measured in the percentage of staff in these positions.
- *Proportion of student from high educated parents (weight 0), Proportion of student from low educated parents (weight .3)* and *Proportion of student from very low educated parents (weight 1.2)* are variables that reflect the student composition of a school. These variables are based on the level of education of one or both of the students' parents. If one or both of the parents attended primary education as highest level a student gets a weight of 1.2. A student gets a weight of 0.3 if one or both of the parents attended the lower tracks of secondary education as highest level of education. Students from higher education parents receive no weights. These variables are measured by the percentage of students in these categories.
- Other variables concerning aspects of the student population are *Number of students from previous Dutch Colonies* and *Number of students from Morocco and Turkey*. These variables represent traditional minority groups of students in Dutch education. Students from Aruba, the Netherlands Antilles, and Suriname are grouped as students from previous Dutch Colonies.

- The variable *percentage of 12 year old students* indicate the amount of grade retention within schools and is measured by the percentage of students of 12 year old at the start of the final year in primary education.
- *One school per board* is a dummy variable indicating whether the school is the only school in a school board or whether multiple schools share the same school board.
- The variable *Denomination* is measured in four categories, namely Public (38.0%), Catholic (24.8%), Protestant (27.6%) and other (9.6%) schools.

Table 7.2b

Descriptive statistics of school characteristics

Variables	*N*	Mean	*SD*
Number of staff	497	22.09	10.97
Staff until 30 years of age	497	20.20	12.26
Staff from 56 years of age	497	4.81	5.59
Percentage of female staff	497	81.64	8.56
Growth of staff	497	3.99	15.01
Staff intake from outside primary education	497	10.32	21.69
Staff intake from other primary school	497	4.38	8.20
Staff leaving outside primary education	497	7.92	18.30
Staff leaving to other primary school	497	3.83	6.53
Percentage of supporting staff	497	10.58	7.57
Percentage of part timers	497	9.47	5.09
Percentage of management	497	50.13	15.31
Growth of number of students	497	1.49	17.55
Proportion of students from high educated parents (weight 0)	500	0.89	0.13
Proportion of students from low educated parents (weight .3)	500	0.06	0.06
Proportion of students from very low educated parents (weight 1.2)	500	0.04	0.10
Number of students from previous Dutch Colonies	500	3.47	9.31
Number of students from Morocco and Turkey	500	10.47	30.95
Percentage of 12 year old students	500	0.02	0.01
Total number of students	500	244.62	133.25
One school per board	500	0.07	0.255

- *Educational vision* is measured in five categories, namely Regular (87.8%), Dalton (2.6%), Jenaplan (4.0%), Montessori (3.2%) and other (2.4%)

- The variable *Urbanisation* indicates whether primary schools are located within or out large cities as measured by the following categories: 4 largest cities of the Netherlands (18.8%), 32 largest cities of the Netherlands (8.4%) and outside the large cities (72.8%).
- *Province* is a categorical variable containing information on the province in which the primary school is located.

7.2.3 Method of analysis

The first step in the process of risk analysis was to estimate the performance of schools based on the test scores of reading comprehension from the Monitoring and Evaluation system. The performance of schools was estimated through the achievement at the end of grade five and the estimated value added over grade three until grade five. Both indicators were estimated using a multilevel growth model with measurement occasions (level 1), nested within students (level 2) and students nested within schools (level 3). The growth models were estimated using the MLwiN 2.25 software (Rasbash et al., 2009). In these multilevel growth models, the scores on the tests from the monitoring and evaluation systems were modeled as a function of time, resulting in an estimation of the average growth of the students in the sample. The time variable in the models was the difference in years between the moment of the test and the end of grade five. Furthermore, multilevel growth models can easily cope with students where data is missing at one or more measurement occasions since a strict balanced design is not necessary. These kinds of models are therefore more flexible than repeated measures analysis (Quené & Van den Bergh, 2004; Snijders et al., 2012).

The final achievement indicator reflects the difference between the performance of students in a particular school at the end of the fifth grade and the average performance of the students in the dataset at the end of grade five. Under-performing schools on the final achievement indicator are those schools with below average performance at the end of grade five. The final achievement indicator can be derived from the multilevel growth model through the school level intercept variance at the end of grade five. Positive intercept residuals indicated above average final achievement of schools at the end of grade five. Value added in these multilevel growth models is the difference between the average growth of students in the dataset and the growth of students in a particular school. Underperforming schools on value added are those schools where the growth of the students is significantly lower than the average growth. The value added indicator can be derived from the growth models

through the school level slope variance of time. Positive slope residuals indicate above average growth over time, while a negative slope residual for a school implies a below average growth over time.

The second part of the analysis is the actual risk analysis. In this risk assessment the "current" performance (*t*), as measured by the valued added and final achievement of schools, was estimated using all "previous" available information of these schools (*t-1, t-2 etc.*). In this case, all information of the schools until the year 2003/2004 (*t-1, t-2 etc.*) was used as predictors of underperformance of schools for the cohort ending in the year 2005/2006 *(t)*. Two statistical methods are investigated for the actual risk assessment, namely linear discriminant analysis and regression tree analysis.

Stepwise linear discriminant analysis (LDA) is a statistical method to find a linear combination of characteristics which separates two or more groups. Discriminant analysis is used when the dependent variable is categorical and multiple independent variables are used as predictors. Discriminant analysis involves estimating a linear equation that predicts in which group a case belongs. In this case whether a school is considered "at risk" or not, given their previous performance and other characteristics. Discrimination analysis is used to test which characteristics contribute most to group separation. For the discriminant analysis missing values for the independent variables were imputated (mode for categorical variables and the mean for continues variables). For each independent variable with imputated missing values a dummy was created indicating missingness. These dummy variables were included in the stepwise discriminant analysis with the other dependent variables.

Regression tree analysis is used in the risk analysis to detect important interactions between possible predictors of the performance of schools (Neville, 1999). Regression tree analysis identifies those characteristics of schools that differentiate most between underperforming schools and schools with average or above average performance. The CHAID algorithm used in this risk analysis finds those differences by using $\chi 2$ tests to measure the association between the dependent variable and the independent variables (Agresti, 1990). The CHAID procedure begins by finding independent variables that have a significant association with the dependent or target variable. It then assesses the category groupings or interval breaks to pick the most significant combination. Categories of the independent variable are combined if they are homogeneous with respect to the dependent variable. The independent variable having the strongest association with the target variable becomes the first branch in a tree with a leaf (also known as a node) for each category that is significantly different with respect to the outcome variable. The process is repeated

for each leaf to find the predictor variable that is most significantly related to the outcome variable, until no significant predictors remain. The subgroups or leaves of data are exhaustive in that they include every data point in the data set and exclusive because each data point belongs to only one leaf. The CHAID algorithm does not exclude missing data. Missing data in regression tree analysis are handled as a separate category, which can be combined with one or more other categories if they are statistically homogeneous.

Both methods of risk models result in a number of characteristics of schools that are associated with underperformance of schools in reading comprehension. Furthermore, the models provide the probability for each school on underperformance. For this paper we considered two rules for classification, namely a probability higher than .50 and a probability of higher than .10 on underperformance as worthwhile for further investigation. The rule of .50 is that standard and .10 is a very conservative rule. The rules will lead to different results in terms of the number of schools that appear worthwhile for further investigation and *false positives* and *false negatives*. A false positive is a school in an end note that shows potential risk-based on the explanatory variables (*until t-1*), while there are no signs of risks on the dependent variable (*at time t*). False negatives are those schools that show risks in the dependent variable while the model doesn't predict potential risks based on the previous performance and characteristics.

To test whether the set of characteristics is robust over time, the rules from the risk analysis of cohort 2005/2006 were applied to the performance of schools from the cohort of 2007/2008. Again, the results are presented in terms of false positives and false negatives. Increasing numbers of false positives and false negatives implies that the model seems not robust over time. This latter indicates whether or not a risk model is useful for the practice of educational accountability.

7.3 Results

7.3.1 Differences in final achievement and value added between primary schools

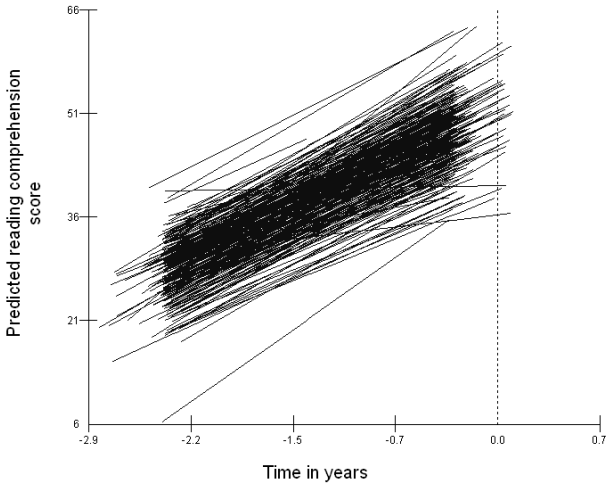
The results of the multilevel growth models for the estimation of value added for the combined cohorts for reading comprehension are presented in Table 7.3. The average score on reading comprehension at the end of grade five of the 2007/2008 cohort is 47.74 points on the latent reading comprehension scale. The final achievement of

students in the two previous combined cohorts seems higher, 51.36 in cohort 2003/2004 and 51.96 in cohort 2005/2006 on the latent reading comprehension scale. Differences in intercepts at the end of grade five indicate that schools differ in terms of their final achievement. For the final achievement at the end of grade five, between 9.2% (cohort 2005/2006) and 11.5% (cohort 2003/2004) of the total variance is accounted for by the school level. These between school differences appear somewhat smaller than previous research into primary schools effects in the Netherlands (Bosker et al., 1997; Verhelst et al., 2003; Wijnstra et al., 2003).

The students from the 2007/2008 cohort grow on average 9.15 points on the reading comprehension scale per year. Similar growth rates were found in the previous combined cohorts. The growth curves of schools for reading comprehension are presented in Figure 7.1. In this figure each school is represented by a line. It can be seen from the figure that most schools take the third grade test at the prescribed moment, however a number of schools take the tests at the start of grade 3. Furthermore, the differences between schools in the performance of students in reading comprehension at the beginning of grade three and the end of grade five are clearly visible. Besides the differences at specific time points, differences in the slopes of the lines indicate differences between schools in growth of reading comprehension. These differences in the slopes of the growth indicate that schools differ in their value added for reading comprehension. In other words, the growth in reading comprehension differs per schools, as in some schools students tend to develop faster than in other schools. For 95% of the schools in the 2007/2008 cohort the growth in reading comprehension ranges between 4.87 and 13.44 points on the latent reading comprehension scale[4]. This implies that in a school with a high value added the growth of the students is over 2.5 times the growth of student in schools with a low estimated value added.

[4] School differences in the growth of reading comprehension are calculated using the following formula: $9.15 \pm 1.96\sqrt{4.87}$. In this formula 9.15 is the average growth of the students in reading comprehension in cohort 2007/2008, and 4.87 is the between school variance for the slope of time.

Figure 7.1
Estimated growth curves for reading comprehension for each school in cohort 2003/2004[5]



[5] Three outliers appear in the growth curve analysis for cohort 2003/2004. One school with very performance at the start of grade three and two schools with very flat lines (very low value added and final achievement). The results of the successive discriminant analysis and regression tree analysis did not appear to change when the scores on the predictors of previous performance (value added, final achievement and underperformance 2003/2004) were set to missing values.

Table 7.3

Multilevel growth models for estimating value added of primary schools for reading comprehension

	Cohort 2003/2004		Cohort 2005/2006		Cohort 2007/2008	
	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.
Fixed Part						
Intercept	51.358	0.382	51.955	0.314	47.741	0.309
Time in years	9.885	0.134	9.907	0.122	9.152	0.130
Cohort	0.669	0.203	0.873	0.186	-0.274	0.162
Random Part						
School level intercept variance	29.779	3.084	22.241	2.219	25.611	2.402
School level slope variance for time	3.764	0.405	3.695	0.366	4.872	0.450
School level covariance intercept and slope	6.351	0.921	4.126	0.706	6.767	0.867
Student level intercept variance	162.103	3.063	155.934	2.734	153.229	2.830
Student level slope variance for time	4.752	0.603	3.950	0.535	2.600	0.576
Student level covariance intercept and slope	9.486	1.071	8.337	0.950	9.736	1.033
Measurement occasion variance	66.617	0.784	64.606	0.705	69.124	0.725
Modelfit						
-2*loglikelihood:	339,340.025		394,705.054		450,852.407	
Number of schools	262		314		371	
Number of students	15,195		17,886		22815	
Number of measurements	43,582		50,921		57863	

7.3.2 Association between final achievement and value added and stability over time

When looking at the between primary school differences at the end of grade five (final achievement) and the growth over time (value added), positive associations were found for all combined cohorts (r = .57; N =371; p < .001; cohort 2007/2008). These positive correlations imply that schools with high final achievement tend to also show more growth over time. In Table 7.4 the schools are grouped based on their observed risks in the cohort of 2007/2008, namely their performance on the value added and final achievement indicators. For this grouping, schools were defined as underperforming on an indicator when they perform significantly below average. Similar, over-performing schools are those schools that perform significantly above average on the performance indicators. For reading comprehension 76 (20.5%) of the 371 schools in the 2007/2008 cohort can be defined as at risk schools, as they show underperformance at one or both of the performance indicators. This grouping of at risk schools is used in the actual risk analysis, to test whether these at risk schools can be accurately identified based on all previous data of these schools. Combining the two performance indicators is based on the relatively high correlation between both indicators and the fact that schools with high initial achievement (start grade three) tend to show less growth over time.

Table 7.4

Distribution of schools of the 2007/2008 cohort over the value added and final achievement indicators

Reading comprehension **N=371**	**Value added: Underperforming**	**Value added: Average**	**Value added: Overperforming**
Final achievement: Underperforming	19 (5.1%)	35 (9.4%)	2 (0.5%)
Final achievement: Average	20 (5.4%)	205 (55.3%)	29 (7.8%)
Final achievement: Overperforming	0 (0.0%)	27 (7.3%)	34 (9.2%)

The past performance of a school might be a powerful predictor of the current performance of a school in a risk analysis. However, this is dependent on the stability of the estimated performance of schools over time. The stability of schools effects is regularly measured by the correlation between the performances of schools over several cohorts. Positive correlations between subsequent cohorts indicate that schools that seem effective for the first cohort tend to perform well for the following

cohort. Correlations between the estimated final achievement and value added indicator over the available combined cohorts are presented in Table 7.5. Mostly moderate positive correlations were found among the performance indicators of primary schools for multiple cohorts. The association between the subsequent cohorts is larger for the final achievement indicator than for the value added indicator. For both final achievement and value added, the association between performance indicators of schools for subsequent cohorts is the strongest and the association decreases in strength for cohorts lying further apart in time.

Table 7.5
Stability of final achievement and value added indicators for reading comprehension over subsequent cohorts

Final achievement		
	Cohort 2003/2004	Cohort 2005/2006
Cohort 2005/2006	.651**	
Cohort 2007/2008	.394**	.483**
Value added		
	Cohort 2003/2004	Cohort 2005/2006
Cohort 2005/2006	.358**	
Cohort 2007/2008	.127*	.342**

**. Correlation is significant at the .01 level (2-tailed).
*. Correlation is significant at the .05 level (2-tailed).

7.3.3 Risk analysis based on Stepwise Linear Discriminant Analysis

A discriminant analysis was conducted to predict whether a school was underperforming in 2005/2006 based on their previous performance (2003/2004) and other school characteristics. The results of this discriminant analysis are presented in Table 7.6. The stepwise discriminant analysis finds seven characteristics of schools that significantly distinguish between average and underperforming schools in 2005/2006. The first predictor variable in the analysis is Percentage students from low and very low educated parents, which is negatively related to the final achievement indicator (r = .41; N =312; p < .001). In schools with a relative high percentage of students from low and very low educated parents, the students reach lower final achievement scores. Final achievement is one of the two indicators that determined underperformance. The second predictor variable from the discriminant analysis is the dummy variable indicating missingness on the indicator Teachers monitor the

progress in the development of students systematically. The third predictor variable is the final achievement of the previous cohort of students (2003/2004). It is not surprising that a previous indication of the performance of schools discriminates between underperforming and average schools in cohort 2005/2006. The fourth predictor variable in the discriminant analysis is the percentage of part timers. The group of underperforming schools tend to have a higher percentage of staff working part time. The last three predictor variables from the analysis are all indicators collected during school inspections. These variables indicate regular evaluation of extra care, sufficient performance of students at the end of primary school and students' engagement with learning activities. Underperforming schools are more frequently judged as insufficient on these indicators. In total, the predictors explain 24.7% of the variance in the grouping variable.

Table 7.6

Results of the discriminant analysis for predicting underperformance in reading comprehension for the 2005/2006 cohort

	Exact F			
Predictor variables	**Statistic**	**df 1**	**df2**	**Sig.**
Percentage students from low and very low educated parents	39.005	1	310	.000
Missingness in indicator Teachers monitor the progress in the development of students systematically	28.301	2	309	.000
Final achievement 2003/2004	22.615	3	308	.000
Part timer ratio	19.497	4	307	.000
Indicator regular evaluation of the effects of extra care	17.211	5	306	.000
Indicator Sufficient performance of students at the end of grade 6 given the student population of a school	15.424	6	305	.000
Indicator Students are engaged with the learning activities	14.234	7	304	.000

Based on the model derived from the discriminant analysis schools are classified as "at risk" schools of the probability of belonging to the underperforming group is over .50 given their background characteristics / risk factors. In Table 7.7 the classifications based on the discriminant analysis and the actual performance are presented. When using a probability of .50 as a rule for classification, 47 schools (15.1%) are classified as at risk according to the discriminant analysis *(t-1)* and are also underperforming *(t)*.

This group are called the true positives. Furthermore, 38 schools (12.2%) show no observed underperformance *(t)*, but according to the discriminant analysis these schools have a high probability on underperformance *(t-1)*. This group is called the false positives. In a risk based educational accountability system, this group of schools would be worthwhile for further investigation. The largest group (119 schools, 63.8%) consists of those schools that show no observed underperformance and no estimated risks. This group of schools are the true negatives. These are the schools do not need any further investigation. A final group is the group op schools that show observed underperformance *(t)*, but that are classified by the discriminant analysis as no estimated risks (28 schools, 9.0%). In a risk based educational accountability system these schools would not get any further investigations as they are not found by the risk model, although they are underperforming. Using the .50 rule implies that 37.3% of the underperforming schools are not classified as "at risk" by the discriminant analysis. This latter group of schools might be of particular importance in risk based educational accountability systems, as one might wish to find all or nearly all underperforming schools.

Table 7.7

False positives and false negatives in cohort 2005/2006

		Cohort 2005/2006	
		Observed performance	
		No	Yes
Estimated risks (prob. 50%)	No	119 (63.8%) *True negatives*	28 (9.0%) *False negatives*
	Yes	38 (12.2%) *False positives*	47 (15.1%) *True positives*
		Cohort 2005/2006	
		Observed performance	
		No	Yes
Estimated risks (prob. 10%)	No	20 (6.4%) *True negatives*	1 (0.3%) *False negatives*
	Yes	217 (76.0%) *False positives*	74 (23.7%) *True positives*

To ensure that more of the underperforming schools are found, more conservative rules can be applied. A similar contingency table is presented in Table 7.7 for the very conservative rule of .10. This means that all schools with a probability higher than .10 of belonging to the group of underperforming schools based on all information until time t-1 are classified as "at risk". Applying this more conservative

rule changes the distribution of the contingency table dramatically. Through applying the more conservative rule more schools are false positives, they will receive further investigation, while there is no actual underperformance. Furthermore, the number of false negatives decreases. This implies that nearly all underperforming schools are found.

7.3.4 Risk analysis based on Regression Tree Analysis

The results of the regression tree analyses for analyzing which set of characteristics relate to underperformance of schools in reading comprehension for the 2005/2006 cohort are presented in Table 7.8 and Figure 7.2. For predicting underperformance in reading comprehension in 2005/2006, the first splitting variable found in the regression tree analysis is the proportion of students from high educated parents within schools. Schools with higher proportions of students from higher educated parents are less frequently underperforming. Based on the proportion of students from high educated parents within schools three groups of schools can be defined. In the first place, schools with equal or less than 73% students from relatively high educated parents (node 1). Of these schools, 64.5% (N=20) are underperforming on one or both performance indicators. The second group is formed by those schools with 73% until 89% students from relatively high educated parents (node 2). In this group 30.5% (N=29) of the schools are underperforming at one or both indicators. The final group contains schools with over 89% students from relatively high educated parents (node 3). This last group has the smallest relative number of underperforming schools (14.4%, N=27). Descriptive statistics show that the proportion of students from high educated parents is mainly related to the estimated final achievement (r = .46; N =314; p < .001) and not with the estimated value added of schools (r = -.01; N =314; p < .842).

The group of schools with over 89% students from high educated parents can be further divided into three subgroups based on their final achievement in the preceding cohort (2003/2004). The group of schools with high final achievement scores in the preceding cohort tend to perform well in the 2005/2006, as only 6.3% of these schools are underperforming (node 6). The second group consists of those schools that were underperforming on final achievement in 2003/2004 (node 5). A large number of these schools is underperforming again in 2005/2006 (54.5%). The final subgroup is the group with average performance on final achievement in 2003/2004 (node 4). Of the schools with over 89% students from high educated parents and

average final achievement in 2003/2004 16.3% (N=16) is underperforming in 2005/2006.

Moving further down the regression tree, it appears that the group of schools with over 89% students from high educated parents and average final achievement in 2003/2004 can be further split up into two subgroups based on the process characteristic of regular evaluation of the effects of extra care. In the subgroup of schools that is considered sufficient in the evaluation of the effects of extra care (node 8), 8.2% (N=5) of the schools are underperforming in 2005/2006. The majority of these schools perform satisfactory in 2005/2006. However, the schools that are considered insufficient in evaluating the effects of extra care (node 7) show more risk on underperformance, as 29.7 % (N=11) of these schools are underperforming in 2005/2006. Both these subgroups can be divided into further groups.

Figure 7.2
Results of the regression tree analysis for predicting underperformance in reading comprehension for the 2005/2006 cohort

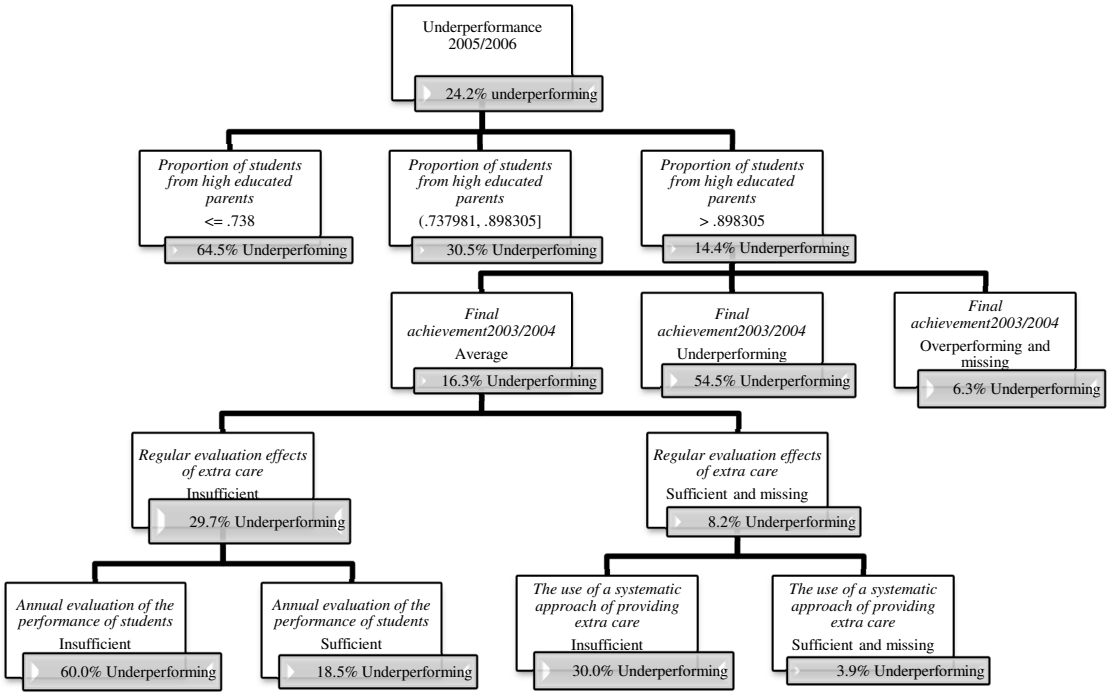


Table 7.8

Results from the regression tree for predicting underperformance on reading comprehension in cohort 2005/2006

Node	Average or overperforming schools		Underperforming schools		Parent Node	Primary Independent Variable				
	N	Percent	N	Percent		Variable	Sig.[a]	Chi-Square	df	Split Values
0	238	75.8%	76	24.2%						
1	11	35.5%	20	64.5%	0	Proportion of students from high educated parents	.000	39.457	2	<= .737981
2	66	69.5%	29	30.5%	0		.000	39.457	2	(.737981, .898305]
3	161	85.6%	27	14.4%	0		.000	39.457	2	> .898305
4	82	83.7%	16	16.3%	3	Final achievement2003/2004	.000	18.894	2	Average
5	5	45.5%	6	54.5%	3		.000	18.894	2	Underperforming
6	74	93.7%	5	6.3%	3		.000	18.894	2	Overperforming and missing
7	26	70.3%	11	29.7%	4	Regular evaluation effects of extra care	.016	7.817	1	Insufficient
8	56	91.8%	5	8.2%	4		.016	7.817	1	Sufficient and missing
9	4	40.0%	6	60.0%	7	Annual evaluation of the performance of students	.014	6.010	1	Insufficient
10	22	81.5%	5	18.5%	7		.014	6.010	1	Sufficient
11	7	70.0%	3	30.0%	8	The use of a systematic approach of providing extra care	.018	7.556	1	Insufficient
12	49	96.1%	2	3.9%	8		.018	7.556	1	Sufficient and missing

The group of schools with over 89% students from high educated parents and average final achievement in 2003/2004, and that are considered insufficient in the regular evaluation of the effects of extra care (node 7) can be divided based on the process indicator annual evaluation of the performance of students. In this group the schools that are considered sufficient in their evaluation of the performance of the students (node 10) are relatively less frequent underperforming in 2005/2006 (18.5%) than schools that are considered insufficient (node 9; 60.0%).

The group of schools with over 89% students from high educated parents and average final achievement in 2003/2004, and that are considered sufficient in the regular evaluation of the effects of extra care (node 8) can be further split up based on the use of a systematic approach of providing extra care. Schools in the subgroup that are considered sufficient in the use of a systematic approach of providing extra care (node 12) are relatively less frequent underperforming in 2005/2006 (3.9%) than schools that are considered insufficient in the use of a systematic approach of providing extra care (node 11; 30.0%).

Table 7.9
False positives and false negatives in cohort 2005/2006

		Cohort 2005/2006	
		Observed performance	
		No	Yes
Estimated risks (prob. 50%)	No	218 (69.4%) *True negatives*	44 (14.0%) *False negatives*
	Yes	20 (6.4%) *False positives*	32 (10.2%) *True positives*
		Cohort 2005/2006	
		Observed performance	
		No	Yes
Estimated risks (prob. 10%)	No	123 (39.2%) *True negatives*	7 (2.2%) *False negatives*
	Yes	115 (36.5%) *False positives*	69 (22.0%) *True positives*

For a risk-based strategy in educational accountability, the end nodes with a relatively large number of schools underperforming are worthwhile for further investigation. In Table 7.9 the classifications from the regression tree models are presented while applying the same rules as for the discriminant analysis. Nodes 1, 5

and 9 have over 50% underperforming schools. When only those end nodes are investigated that show over 50% underperforming schools 52 schools would appear worthwhile for further investigation. Less than half of the underperforming schools are found using this classification rule (32 schools, 42.1%).

For the cohort of 2005/2006 this includes nodes 1, 2, 5, 9, 10 and 11, when applying the .10 rule. A further investigation of six end notes implies that there is no single set of characteristics that identify underperforming schools for reading comprehension. These six end nodes with a relatively large at risk schools include in total 184 primary schools; this is 59% of the total sample. These nodes contain 69 of the 76 underperforming primary schools in reading comprehension for the cohort 2005/2006. Seven underperforming primary schools will not be found in these nodes *(false negatives)*. Furthermore, 115 primary schools in these end nodes are not underperforming in cohort 2005/2006 and they can be called *false positives*, as these schools will be further investigated by inspectors while there is no actual underperformance.

When comparing the classifications of the risk models from the discrimination analysis and the regression tree analysis, it appears that the predictors that arise in both models are fairly similar. Furthermore, the regression tree approach shows more false negatives, schools with observed underperformance *(t)*, that are not found by the models *(t-1)*. For the .10 rule there is a difference of 6 false negatives. However, the regression tree approach leads to a smaller number of false positives, schools that show no actual risk, but based on the models they appear worthwhile for further investigation. When applying the .10 classification rule, the discrimination analysis leads 102 more false positives.

7.3.5 Applying the risk models on the performance data of cohort 2007/2008

The robustness of the models is investigated by applying the splitting rules based on the previous discrimination analysis and the regression tree analysis of cohort 2005/2006 on the performance data of the cohort of 2007/2008. Results of applying the models on the performance data of cohort 2007/2008 in terms of false and true positives and negatives are presented in Table 7.10. For this robustness check we only applied the .10 classification rule. Although the models show differences in the results for the 2005/2006 cohort, the results for the 2007/2008 cohort, in terms of true and false positives and negatives are rather similar.

Applying the discriminant function to the performance data of the following cohort leads to 219 schools that appear worthwhile for further investigation. This is 59% of the total sample. Compared to the results of the discrimination analysis of the 2005/2006 cohort there seems a drop in the percentage of schools that appear worthwhile for investigation, 93% in 2005/2006 and 59% in cohort 2007/2008. Furthermore, based on the discriminant analysis 80% (61 of the 76) of the underperforming schools are found in the 2007/2008 cohort. This implies that applying the discriminant function to the following cohort leads to 15 false negatives, schools that are underperforming but that are not found. Compared to 2005/2006 this is an increase of the number of false negatives.

Table 7.10
False positives and false negatives in cohort 2007/2008

Discrimination analysis		**Cohort 2007/2008**	
		Observed performance	
		No	Yes
Estimated risks (prob. 10%)	No	137 (36.9%) *True negatives*	15 (4.0%) *False negatives*
	Yes	158 (42.6%) *False positives*	61 (16.4%) *True positives*
Regression tree analysis		**Cohort 2007/2008**	
		Observed performance	
		Yes	No
Estimated risks (prob. 10%)	Yes	139 (37.5%) *True negatives*	14 (3.8%) *False negatives*
	No	156 (42.0%) *False positives*	62 (16.7%) *True positives*

Based on the previous regression tree model 218 schools of cohort 2007/2008 appear worthwhile for further investigation, which is 59% of the sample. This implies that applying the regression tree model to the data of cohort 2007/2008 will not lead to an increase of the relative amount of schools that are worthwhile for investigation. Based on these rules 81.6% (62 of the 76) of the underperforming primary schools in the 2007/2008 cohort are found. Applying the rules from the regression tree model on the data of a subsequent cohort will lead to 14 false negatives, schools that are underperforming in 2007/2008 but that are not found in the six end nodes of interest.

Applying the model on data of a subsequent cohort will therefore lead to an increase of the number of false negatives; 7 in cohort 2005/2006 and 14 in cohort 2007/2008. Furthermore, the model leads to 156 false positives for the 2007/2008 cohort.

7.4 Conclusion and discussion

The purpose of the current study was to determine which characteristics of schools are relevant for predicting "at risk" schools in primary education and to assess the robustness of a risk model over multiple cohorts. Within the context of educational accountability, low performance of students within schools at the end of a formal stage of education and a little growth in performance of their students during this formal stage of education were defined as risks. In this paper, we therefore defined a school "at risk" as a school with low final academic achievement and/or low value added. Two statistical models have been applied to predict which schools are "at risk", namely a discriminant analysis and regression tree analysis. Taken together, the composition of the school in terms of the proportion of students from high educated parents, previous performance of schools, evaluation of effects of extra care, monitoring the performance of students and the use of a systematic approach in providing extra care appear as the best predictors of underperformance of the primary schools in the sample. However, the results of the regression tree analysis indicate that "at risk" schools cannot be described by a uniform set of characteristics. For reading comprehension, the regression tree analysis for the cohort 2005/2006 showed multiple end nodes with high risks on underperformance.

The results of both models indicate that, if risk models for predicting underperformance of primary schools would be applied in the context of educational accountability, a large number of schools need further investigation to find nearly all underperforming schools. For example based on the regression tree analysis 59% of the schools need further investigation to find 69 of the 76 underperforming primary schools. Based on the discriminant analysis even more schools need further investigation. If the rules from the both statistical models are applied on performance data of schools for a later cohort (cohort 2007/2008), the number of false negatives increases, schools that are underperforming in 2007/2008 but are not found in the end nodes of interest.

Risk-based strategies in accountability are considered more efficient and effective than traditional strategies (Sparrow, 2000). An efficient risk model for educational

accountability would need to find nearly all underperforming schools, thus a very small number of false negatives, with the smallest possible number of false positives (satisfactory performing schools that need further investigation). However, the efficiency of a risk-based strategy depends heavily on an accurate prediction of future risks. In the current study, only moderate correlations were found between the estimated performance indicators of primary schools between subsequent cohorts. Differences in the performance of schools between subsequent cohorts might provide an explanation for the large amount of false positives in the models. All in all, results from this risk analysis imply that underperforming schools cannot be predicted very accurate. However, there appears a group of about 40% of the schools that show very small risk on underperformance. These schools represent the efficiency gain when these models would be applied in a risk based accountability strategy. Although, one might hope for more accurate classifications of underperforming schools, these models provide some information to realize a more efficient accountability strategy.

The number of false positives and false negatives in a risk analysis depends heavily on the decision rules that determine which schools are worthwhile for further investigation. In this study, the rules were that end nodes with over 10% and 50% underperforming schools or schools with a probability of .10 and .50 based on the discriminant analysis should be further investigated. Both models show large differences in the classifications of schools for both rules. When applying the rule of 10% the number of false negatives appeared relatively small, while the number of false positives became relatively large. With these rules, one would need to investigate a large number of schools and therefore the model can be considered inefficient, however, almost all at risk school will be found. Through applying less conservative rules less schools need further investigation at the expense of an increase in the number of false negatives, schools that are underperforming but that are not found in the end nodes of interest. Therefore less conservative rules attenuate the effectiveness of the risk model in identifying possible underperforming primary schools. Furthermore, in the 2005/2006 cohort differences were found between the two statistical models. These differences might arise from the models; combinations of linear effects in the discriminant analysis and complex interactions in the regression tree analysis. Besides the effects of the type of statistical models, the way of handling of missing values might account for a part of the differences in classifications.

When interpreting the results it should be kept in mind that the sample of schools in this study appeared relatively homogeneous and not representative for the population of primary schools in the Netherlands. In this sample, 3.7% of the schools were considered inadequate and 0.4% very poorly, when the inspection judgments in

2009/2010 are considered. In the population of primary schools in the Netherlands 6% of the schools were considered inadequate and 1.5% of the school very poorly at January 1st 2010 (Inspectie van het Onderwijs, 2012). Similarly, the dataset contained schools with a relatively high proportion of students from highly educated parents. Despite this homogeneous sample, considerable differences were still found between schools in their final achievement and value added. However, this homogeneous sample might lead to some bias with respect to the characteristics associated with underperformance. Therefore, results of this study cannot be applied directly in educational accountability, but do provide an example of risk analysis.

In this article we defined underperforming schools as those schools with insufficient final achievement and/or insufficient value added. The choice of risk in a risk analysis determines the results of a risk analysis to a large extent. Differences can be expected in 1) the choice of statistical model, 2) characteristics of schools that predict underperformance, 2) classifications and miss classifications. Combined these three factors determine whether applying a risk based strategy will lead to considerable efficiency gains.

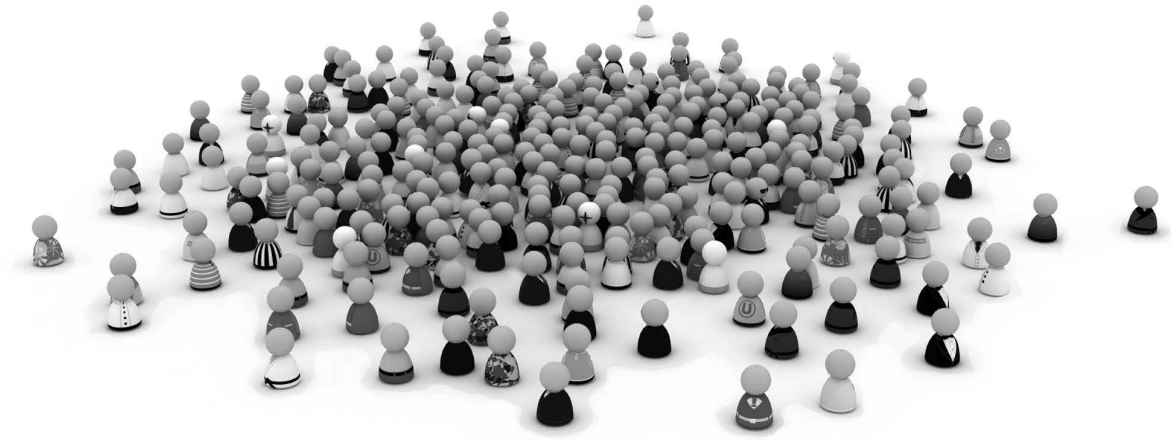
Moreover, the data in this study are derived from the CITO Monitoring and Evaluation system. The tests on which the performance indicators of schools are based are administered by the schools to monitor the performance of their students. This implies that administering conditions might differ between schools and between several tests on the same school. Furthermore, schools might differ in selection of students included in the tests and the number of tests administered each year. While most schools test their students annually, some schools test their students half way and at the end of each school year. For the latter schools more data is available for the estimation of value added and final achievement. When some schools test their weakest students more often than average or above average students this might bias the estimated final achievement and value added.

Furthermore, it should be noted, though, that in the estimation of value added only the test scores of students were used in the analysis to predict the performance of schools, as indicators of the ethnic and socio-economic background of the students were not available in the student level dataset. Moreover, the sample of schools in this study is relatively small (n = 371), which leads to a small power in the regression tree and discrimination analysis given the large number of possible predictors of underperformance of primary schools.

The consistency of the results of risk analyses over multiple studies is considered as an important requirement before risk models can be applied in practice (Gibb,

1997). Previous studies within the tradition of educational effectiveness research have merely focused on identifying characteristics of effective schools instead of underperforming or failing schools (Reynolds et al., 1999a). Although sets of characteristics of effective schools have been formulated over time, comparing the results of individual studies show considerable inconsistencies in the characteristics of effective schools (Teddlie et al., 2000a). Based on this previous research one might expect that, in time, multiple studies into underperforming schools might lead to some inconsistencies in the results of associated characteristics. Furthermore, this indicates that risk models cannot yet be adequately applied in the practice of educational accountability. Future research is needed to show which sets of characteristics are consistently related to underperformance of schools.

Chapter 8
General Discussion



The main aims of the studies conducted in this dissertation were a) to develop value added models for school within the context of educational accountability, b) to study whether or not using value added in educational accountability would lead to valid comparisons of the performance of schools, and c) whether value added can be used in a risk based educational accountability system to predict future underperformance of schools. The studies in this dissertation focussed on primary, secondary and senior secondary vocational education in the Netherlands. In each study, a specific aspect of value added was studied in more detail to test whether value added models could be developed and to get some indications to what extent value added is useful for educational accountability and which weaknesses should be acknowledged. Furthermore, the studies discuss possibilities how to deal with the weaknesses of value added. This final chapter of this dissertation will summarize the main findings from the previous empirical studies. After that, theoretical and practical implications of the results of this will be presented in paragraph 8.2. Paragraph 8.3 will provide an overview of limitations of the study conducted in this dissertation. Finally, this dissertation will close with directions for future research.

8.1 Main findings

The conceptual and empirical differences among various value added models in secondary education are described in the second chapter of this dissertation. Value added indicators in educational accountability systems have many different operationalizations although they are used for the same purpose. These differences between the value added indicators can be found in the statistical modeling (status or growth models) and in the sets of covariates used to control for factors outside the school. Using different sets of covariates means estimating different schools effects (for example type A, B or X school effects) with different conceptual meanings (Willms et al., 1989; Raudenbush et al., 1995). The models used in educational accountability vary in their use of covariates between only prior achievement in the Tennessee Value Added Assessment System (Sanders et al., 1994) to an extensive amount of control variables including prior achievement, student level characteristics and school composition in the Contextual Value Added indicator (Ray, 2006). Modelling various value added models with different sets of covariates on data from

the VOCL'99 cohort shows moderate to high correlations between the estimated school effects from the models, which means that the different models seem to measure a similar construct. However, the in- or exclusion of covariates changed the classification of school as underperforming, average or overperforming. A misclassification can have important implications for individual schools, such as intensified inspections. The association (Kappa) between the models varying in covariates in the classification which schools were underperforming, average or overperforming varied from fairly to almost perfect.

The study in chapter three investigated the effects of modelling student mobility and long term effects of primary schools on value added estimates of secondary schools. Both student mobility and long term effects of primary education are examples of imperfect hierarchical structures in educational data (Snijders et al., 1999; Goldstein et al., 2007; Leckie, 2009). Student mobility, that is students changing schools during a formal period of education, leads to the membership of these students to multiple schools, in this case secondary schools. In case of long term effects of primary schools, students are both nested within a secondary school and a primary school. Primary as well as secondary schools are populations of interest in the so-called cross classification models. However, in traditional value added models students are usually assigned to the school where they made their final examination, regardless whether they changed schools or not. Empirical analysis of data from the VOCL'99 cohort shows that for the VMBO theoretical track, value added estimates are biased if students are only assigned to their final school. Although the correlation between the multiple membership model, incorporating student mobility, and a traditional model appeared high, the impact for individual schools is again rather high, especially for schools around the average. Shifts in rank order of ten or more places were found for over 50% of the schools in the sample. The most and least effective schools are rather stable over both models. Incorporating long term effects of primary schools did not appear to change the estimated value added for secondary schools, at least for this sample.

In the fourth chapter, effects of schools on both cognitive and non-cognitive outcomes in Dutch secondary education, in the context of educational accountability, were explored by means of multivariate multilevel analysis. The sample for this study consisted of 10,849 students in 82 schools from the VOCL'99 cohort. Four dependent variables were considered in this study, namely the performance of students in Dutch language, mathematics, perceived classroom climate and achievement motivation. Our study confirms that the relative influence of schools is higher for the cognitive domain than for the non-cognitive domain. Moderately strong

correlations were found between school effects on the perceived classroom climate and school effects on mathematics and language achievement. Schools that perform well on language tend to perform well on the perceived classroom climate. These moderate positive correlations imply that effectiveness of schools in the cognitive domain doesn't necessarily have to be detrimental for outcomes in the non-cognitive domain. In this sense, the results support the multidimensionality of educational effectiveness as advocated in previous studies (Thomas et al., 2000; Gray, 2004b).

Estimating added value as an indicator of school effectiveness in the context of educational accountability is usually based on test or examination score. The fifth chapter of this dissertation investigated the possibilities of using indices of educational careers instead. A number of advantages of a value added indicator based on educational careers of students can be formulated, such as: (a) the societal significance of the educational career of students for stakeholders, (b) the fact that a single indicator can be estimated for an entire school in differentiated educational systems, where not all schools provide education in all tracks, and (c) the expectation that value added based on educational careers is more robust against strategic behaviour. Excluding the weakest student from taking tests and test manipulation is no longer an option if administrative data is used on educational careers. Furthermore, with respect to reshaping the test pool, value added based on test scores and value added based on educational careers leads to opposite incentives for schools. Empirical analysis of Dutch cohort data (VOCL'99) for secondary education showed considerable differences in effectiveness between schools in the careers of students. After controlling for student differences and school composition 7.1% of the variance in educational careers is associated with the school level. Furthermore, differential school effects were found for both socio-economic status and prior achievement. The phenomena of differential school effects for socio-economic status and prior achievement are linked to differences between schools in the tracks in which the schools provide education. Schools that provide education in the general tracks (VWO and HAVO) tend to have steeper slopes for prior achievement and socio-economic status. Secondary schools with only pre-vocational education (VMBO) tend to have more flat slopes. Students from less affluent families and low prior achievement when entering the secondary school benefit more from schools with relative flat slopes.

In Chapter 6 the possibilities of estimating value added as a performance indicator in senior secondary vocational education were investigated. Although value added indicators of both primary and secondary education have been developed since the 1980s, the research on school effectiveness has largely neglected vocational education because of its complexity. Studies on the development and methodology of

value added indicators in the segment of senior secondary vocational education (Harmon, 1992; Armstrong et al., 2000) or higher education (Yunker, 2005; Rodgers, 2007) are scarce. The great variety of its student population and its complex structure including multiple training programmes for different degrees, make senior secondary vocational education a challenging sector of education in terms of developing quality indicators. For estimating value added indicators data of almost 90,000 students in the Dutch senior secondary vocational education are used. Factors such as ethnicity, living in problematic neighbourhoods, and students' prior educational attainment appear to be significant predictors of the outcomes of senior secondary vocational education. The results indicate a considerable variance in the effectiveness among clusters of training programmes, whereas among large educational institutions this is close to zero. Of the total variance among the student outcomes 14% refers to the training programme clusters after controlling for student characteristics. About 30% of the training programmes studied can be identified as over-performing and 20% as underperforming. Furthermore, the predicted probabilities of obtaining a diploma differ considerably between underperforming and over-performing training programmes. Compared to the unadjusted scores 20% of the clusters of training programmes are given different classifications in terms of underperforming, average, and over-performing, whenever a value added model is used.

The final empirical study in this dissertation (Chapter 7) explored the possibilities of empirical risk modelling in the context of educational accountability. A risk based approach, in which the intensity and/or frequency of school inspections vary across schools dependent on the risk level of a specific school is a recent development in educational accountability. A risk-based inspection system is considered to be more effective because it enables inspectorates to focus on organizations at risk. In this chapter we assessed which characteristics of primary schools are relevant in predicting which schools are "at risk" and how robust a risk model is over multiple cohorts. The empirical risk model was based on a sample of 500 Dutch primary schools. At risk schools were defined as schools performing significantly below average on the final achievement and/or value added indicators. Two statistical methods were explored to predict underperformance. The composition of the school in terms of the proportion of students from high educated parents, previous performance of schools, a systematic approach and evaluation of effects of extra care and monitoring the performance of students appear as the best predictors of underperformance of the primary schools in the sample. The results from both statistical methods indicate that, if risk models for predicting underperformance of primary schools would be applied in the context of educational accountability, a large number of schools need further

investigation to find nearly all underperforming schools. By applying less conservative rules fewer schools need further investigation at the expense of an increase in the number of false negatives, that is, schools that are underperforming but that are not found. Therefore less conservative rules attenuate the effectiveness of the risk model in identifying possible underperforming primary schools.

8.2 Theoretical and practical implications

The studies in this dissertation have several practical implications for further development and implementations of value added in educational accountability. This paragraph starts with general implications that arise from multiple studies in this dissertation and thereafter, implications from the separate chapters are discussed.

In the first place, throughout all of the studies the value added indicator shows relatively small differences between schools (for all educational sectors) and relatively large uncertainty, resulting in large standard errors surrounding the estimated effect of a school. This is clearly visible in the study into value added in senior secondary vocational education, but it holds also for the other sectors of education. These small differences between schools and large uncertainty cause a low distinctive character of the value added indicator in general. These findings correspond to findings from previous studies into value added in the United Kingdom (Goldstein et al., 1996). This has important implications for the way the indicator can be used in educational accountability. At best three groups can be distinguished. The largest of the three groups are the schools for which the estimated value added is not significantly different from average. Two smaller groups can be identified with estimated effects significantly different from the average, namely the under- and overperforming schools. It is for this same reason that Goldstein (1997) described value added as a crude screening device at best. However, these small differences between schools and large uncertainty are not only a drawback of value added, but for most indicators of educational quality. This holds also for performance indicators currently used in educational accountability. In primary education this problem is more prominent as the sample sizes per primary school used to estimate value added are usually smaller than the sample sizes in secondary and vocational education. Combining data from two or more successive cohorts, as applied in the risk analysis (chapter 7), might be a practical solution to this problem of uncertainty. Using information from two or more cohorts increases the sample size per school and therefore more information is available to make a more reliable estimation of the value added of a school. However,

this method of combining data from successive cohorts relies on the relative stability of school effects.

Furthermore, three types of data have been used in this dissertation, namely cohort data (VOCL'99) in secondary education, National student data (BRON) from senior secondary vocational education and data from the monitoring and evaluation system (LOVS) in primary education. Results from the various empirical chapters show that estimating value added is possible for these sectors of education, when data at the student level is available in which at least a measure of prior and final achievement is recorded. The three types of data applied in the empirical studies in this dissertation differ in the possibilities to estimate value added. The most important differences arise from the availability of control variables and test scores. The cohort data used in the studies for secondary education contains many possible control variables and several measures of prior achievement. This type of data is in particular appropriate for estimating value added based on status models, the final achievement of students controlled for prior achievement and other control variables. Value added derived from such models can be interpreted as a measure of relative *achievement* of students in one school compared to students in other schools in the same sample. These models only implicitly measure the growth of students. Similar types of value added models can be estimated on the National Datasets. However, data from the monitoring and evaluations systems are more appropriate for estimating value added based on growth models. This data contains information on the performance of students on multiple occasions and measured on the same latent scales. Estimates of value added derived from these type of model have a slightly different interpretation, namely a measure of relative *progress* of students in one school compared to students in other schools in the same sample. The growth of students during a particular period is explicitly modelled. Nevertheless, background characteristics of students were only available for a very small percentage of students. This implies that the value added models based on this data only control for previous achievement of students and that it remains questionable whether differences in growth rates between schools are partly due to differences in student population and other background characteristics. Based on the data collection in the three sources of data differences arise in the representativeness of the samples of students and schools. For example, participation in the cohort study is voluntary and for the data from the monitoring and evaluations system only schools are included that use this system for a relatively large number of their students.

In chapter two empirical and conceptual differences among various value added models are discussed. The results indicate that value added models derived from

statistical models with different sets of covariates lead to different classifications of individual schools as underperforming, average or overperforming. These differences stress the importance of a deliberate and widely accepted choice which covariates should be included in the value added model. The following points may be considered while making this choice, 1) conceptual differences, 2) empirical differences in classifications, and 3) the implications of the inclusion of covariates for subgroups of students.

The empirical study in chapter 3 showed that ignoring student mobility might lead to changes in the estimated value added of individual secondary schools. However, the complexity of the statistical model to allow for student mobility conflicts with the need for simple and transparent indicators in educational accountability. Furthermore, the restrictions on the data posed by the multilevel multiple membership models make it unusable for educational accountability in a differentiated educational system. The bias caused by ignoring student mobility seems to have influenced the value added estimations the most for average schools. The estimated value added of underperforming schools, that are of particular interest in educational accountability, did not seem to change due the in- or exclusion of student mobility in the statistical modelling of value added. This last finding implies that the influence of ignoring student mobility in the estimation of value added will not lead to very different identifications of underperforming schools. However, only three studies have been reported yet in which the effects of multiple membership models on the estimated value added are discussed. More research is needed to investigate the effects of student mobility on performance indicators in educational accountability before a deliberate choice can be made to ignore student mobility in the estimation of performance indicators for the particular use in educational accountability.

Results from the study presented in chapter 4 show that the variance between schools is considerably smaller for the non-cognitive outcomes. This finding can be explained through the fact that the cognitive domain is explicitly taught in schools. Non-cognitive outcomes are usually a more implicit part of a schools' curriculum (Dijkstra et al., 2001; Peschar, 2004). Furthermore, the moderate positive association between the effectiveness of schools in the cognitive and non-cognitive domain implies that a focus on the cognitive domain doesn't necessarily have to be detrimental for outcomes in the non-cognitive domain. The outcomes used in this study are not a definite set of possible cognitive and non-cognitive outcomes. More research should be conducted to investigate between school differences and the association between indicators for multiple outcomes. Based on this and future results a choice should be made whether non-cognitive outcomes should be included in

educational accountability systems to get a more detailed image of a schools' performance. The following points should be considered while making this choice. Performance indicators in educational accountability are only relevant if there is substantial variance between schools. A consequence of little variance between secondary schools and large confidence intervals for the estimated school effects, usually found for the non-cognitive domain, is that indicators can poorly distinguish between underperforming, average and overperforming schools. If there is a satisfactory amount of between school variance one might wish to include multiple outcomes, however in accountability systems with multiple indicators of the performance of schools it gets more complicated to establish which schools are performing well or not.

In chapter 5, a value added model is explored based on the educational careers of students. Results indicate considerable differences between secondary schools in promoting educational careers for their students. Furthermore, differential school effects were found for prior achievement and socio-economic status. An implications of these results is that no single indicator can be used to for the effectiveness schools, but multiple indicators for subgroups of students seem necessary for an accurate and detailed picture of the effectiveness of schools on the students educational careers. Although it can be argued that value added based on educational careers is more robust against some forms of strategic behaviour, it might be considered to use value added based on educational careers and test or examination scores as complementary indicators in educational accountability. Using multiple indicators might be beneficial for the robustness of an accountability system with respect to strategic behaviour, as multiple indicators usually provide opposite incentives for schools (Koretz, 2003).

In this dissertation value added has been estimated for clusters of training programmes and educational institutions in senior secondary vocational education using multilevel techniques (chapter 6). Research in educational effectiveness research has largely neglected the vocational education sector and existing studies regularly used single level statistical models for their estimation of value added. The results and methodology of estimating value added in senior secondary education provide a starting point for future research and development for educational accountability for this educational sector. However, the results of this study clearly indicate that more variance in student performance is associated with training programmes than with educational institutions. This indicates that the large educational institutions are fairly similar in their performance, but within educational institutions differences in effectiveness arise between clusters of training programmes. Current performance indicators in Dutch educational accountability for senior secondary vocational

education are measured at the level of educational institutions. Based on the results of the current study one may question if the level of educational institutions in the right level of inference.

In chapter 7, an exploration into risk analysis based in value added was conducted. The most important finding of this study is the fact that many schools need further investigation to find nearly all underperforming schools. This implies that underperformance of schools in the future cannot be estimated very accurately. The risk model in itself is therefore not very efficient. This has major implications for the use in educational accountability. Given that many schools need further investigation a risk based accountability strategy moderate efficiency gain.

8.3 Limitations

Concerning the subject value added as indicator of school effectiveness in the context of educational accountability, many aspects of validity, reliability and usability of value added can be investigated. This dissertation focusses on a restricted number of aspects of the validity of value added in several educational sectors. Therefore, this dissertation will give a limited picture of the validity of value added.

The validity of value added is one of the most important requirements for the use of such an indicator in educational accountability systems. However, establishing which value added indicator is the most valid cannot directly be answered through an empirical analysis in which multiple models are compared. A comparison of value added estimates of schools from two statistical models will not answer the question which of the value added estimates is the closest to the true school effect. Furthermore, comparing multiple value added estimates of schools from statistical models leads to the following question: "When can two estimates derived from different statistical models be considered as similar or strongly associated and when should one conclude that two value added estimates are different?" Differences between value added estimates can be described in terms of changes in ranks, miss-classifications and the association between estimates can be described as correlations or agreement. There is no easy line or rule in the percentage of miss classifications, changes in ranks or correlations upon which conclusions on similarity or differences between several value added estimates can be drawn. Especially in the context of educational accountability, where miss classifications can have important consequences for individual schools, such as intensified inspections, the question becomes even more prominent.

With the exception of senior secondary vocational education, large scale national datasets for estimating value added were not yet available in the Netherlands. As an alternative option we have used cohort data for the estimation of value added in other educational sectors. This has important consequences for the generalizability of the results of the studies in this dissertation to the context of educational accountability in the Netherlands. Cohort data contains usually more detailed information on the performance and characteristics of students, which might lead to differences in the estimated value added for schools and the subsequent validity of the estimated school effects (Timmermans et al., 2011). For Dutch secondary education, student level data is already available, although, at this moment, it lacks valid measures of prior achievement of the students. Therefore, these data does not yet allow an adequate modeling of value added, as an indication of prior achievement is considered essential for a valid estimation of the effects of schools (Willms, 1992). The first cohorts of students in secondary education for whom the school leavers test and the primary school teachers advice are recorded in the national student datasets will take their final examinations within the near future. This offers the possibility for a further investigation of the long term primary school effects and student mobility on the estimates of value added, other validity aspects and the usefulness of value added for purposes. In Dutch primary education, national student level data sets are under development. Furthermore, through recent policy changes the school leavers test at the end of primary education will probably become mandatory in the future. However, whether value added can be estimated for Dutch primary schools on these national datasets depends on the availability of measures of the performance of the students at an earlier stage during primary education in the future.

Missing values on one or more control variables were a regular phenomenon in the studies conducted in this dissertation. In all three types of data missing values were found to a lesser or greater extent. Missing values in the control variables impose important consequences for the validity of value added, because the least able candidates are the most likely to be excluded due to missing data (Rubin et al., 2004). And a method which includes only the complete cases will give an upward bias in the estimates of school and pupil performance (Thomas et al., 1997b). Furthermore, results of imputation strategies may lead to biased results as the missingness is mostly not at random. Differences in the amount of missingness might arise from the method of data collection between the different types of datasets. In the national datasets, for secondary and vocational education, the background characteristics are recorded for (almost) all students. However, missing values remain problematic, because in this type of data the availability of prior achievement poses a source of

possible missing values. For example, there will remain students that enter training programmes in vocational education from the professional life for whom an indication of prior achievement cannot be established.

The particular focus of this dissertation on the usefulness of value added for educational accountability excludes other possible purposes of the indicator. Examples of other possible uses of value added are school choice by parents, data driven teaching by teachers or school teams, and for school boards to discuss school performance with stakeholders and inspectors. Each use of value added comes with its own requirements for the indicator. Parents, for example, might like to know at which school their child is expected to perform best, regardless whether this performance is caused by the school or the student population at that school (Willms et al., 1989; Raudenbush et al., 1995). For accountability, however, it is very important to isolate the effects of the school from the effects of the student population as good as possible. The instability of school effects is a major problem in the usability of value added for school choice since there is a gap of 5 until 7 years between the cohorts used to estimate value added for the league tables and the cohort of your own child graduating (Leckie et al., 2009; Leckie & Goldstein, 2011a; Leckie & Goldstein, 2011b). Instability of schools effects in educational accountability leads to less predictive power in risk assessments and therefore a less efficient model. However, when value added estimates are available each year the instability imposes a smaller problem than for school choice. Moreover, for data driven teaching, schools are interested in detailed data on subjects, tests, individual or subgroups of students, while inspectorates in the context of educational accountability search for as few indicators as possible to reflect the performance of the schools. This is why results from this dissertation cannot be generalized to other possible applications of value added straight-forwardly.

8.4 Future directions

One of the studies in this dissertation describes the possible method of modelling value added in senior secondary vocational education. The educational effectiveness research tradition has focused largely on primary and secondary education and only a few studies in estimating school effects in other educational sectors have been undertaken (Harmon, 1992; Armstrong et al., 2000). The Dutch inspection framework for senior secondary education is limited also in indicators of educational quality (Inspectie van het Onderwijs, 2011f). Before value added, as educational quality indicator, can be used in educational accountability systems it should be taken through

a rigorous validation process. The estimation method described in chapter 6 of this dissertation might be a starting point in the validation of estimating value added and educational effectiveness research in higher or senior secondary vocational education. Basic concepts from educational effectiveness research and indicators of validity, such as differential school effects, stability over time and consistency over outcome measures should be investigated for higher and vocational education. Furthermore, it might be worthwhile to investigate the association with existing quality indicators in vocational education and to discuss the results of value added judgments with experts in the field.

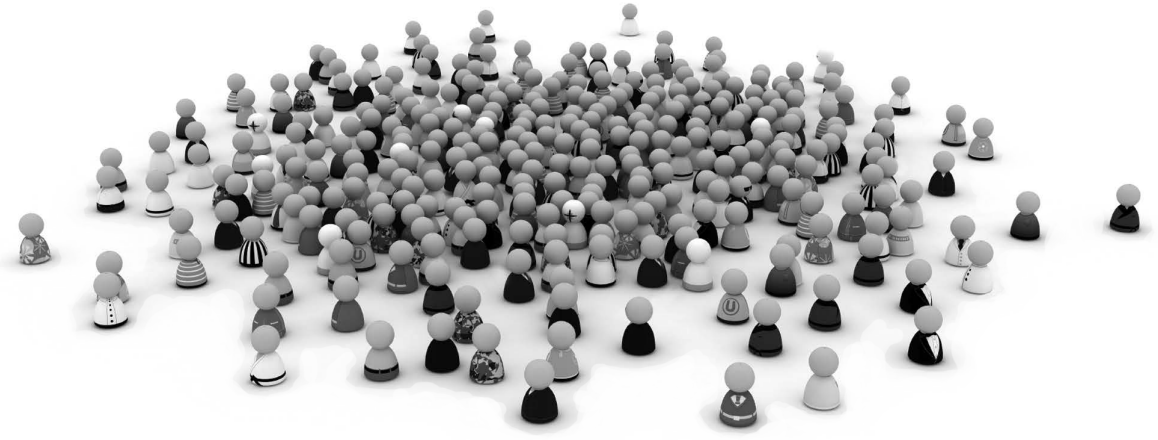
The studies in this dissertation have focussed merely on the validity and the possibilities of estimating value added in the context of educational accountability. Other research on indicators of educational effectiveness shows a similar focus, with specific attention for complex statistical models and validity issues. However, the aspect of which information is necessary for inspectors, school boards and teachers to interpret these performance indicators correctly remains a subject of future research. First steps in research have been taken in which performance indicators are presented in alternative ways to illustrate the uncertainty of the indicators (Leckie & Goldstein, 2011b).

Chapter 5 of this dissertation focussed on value added based on the educational ladder, as an example of an efficiency indicator. The results of this study suggest that a single indicator based on all students within schools have several problems for the context of educational accountability. Both differential school effects and differences in value added between schools with a different structure indicate that a single value added indicator is insufficient in providing valid indications of the efficiency of a secondary school. This implies that further work needs to be done in a) developing indicators for specific subgroups of students, based on their prior achievement and socio economic status, b) testing whether indicators for subgroups of students are comparable over schools with a different structure c) investigating the possible association between short term en long term secondary school effects.

Risk based strategies are a recent development in educational accountability. Although risk models are widely used in other disciplines such as medicine and economics, these models have not yet been applied to educational data. The empirical study in chapter seven provides a first exploration in risk modelling in education. However, a lot of work needs to be done in testing the validity, reliability and robustness of such risk models in education. Future research could therefore focus on a) analysing larger scale empirical risk models, b) the consistency of characteristics of

underperforming schools over multiple studies, c) comparisons of various statistical models to estimate potential risks of units, d) the reliability of the estimated risks, e) and differences between risk models if different outcome measures are considered.

Nederlandse Samenvatting



In de afgelopen decennia zijn in verschillende landen toegevoegde waarde indicatoren opgenomen in het onderwijstoezicht, om dat deze indicatoren de belofte in zich dragen een eerlijke vergelijking mogelijk te maken tussen de prestaties van verschillende scholen. Voorbeelden van toegevoegde waarde indicatoren die op dit moment gebruikt worden in onderwijstoezicht zijn Contextualized Value Added (Ray, 2006; Ofsted, 2010), Tennessee Value Added Assessment System (Sanders et al., 1994; Sanders, 2003) en het Colorado growth curve model (Betebenner, 2007; Betebenner, 2009).

Toegevoegde waarde is een van oorsprong economisch concept gebaseerd op de input, de processen en energie en de output van organisaties of bedrijven. Toegevoegde waarde is geïntroduceerd als een indicator voor de kwaliteit van scholen. In de loop van de tijd zijn er verschillende definities gegeven van het concept. Een voorbeeld is "de bijdrage van een school op de ontwikkeling van leerlingen in de richting van een voorgeschreven doel (bijvoorbeeld cognitieve prestaties). Deze bijdrage van scholen is netto van andere factoren die bijdragen aan de ontwikkeling van leerlingen" (OECD, 2008, p. 17). Een andere definitie van toegevoegde waarde is "een indicatie van de mate waarin een school bijdraagt aan de ontwikkeling van alle leerlingen in een verscheidenheid van vakken in een gegeven periode in vergelijking tot andere scholen in dezelfde steekproef" (Sammons et al., 1997). In dit proefschrift hanteren we een bijna vergelijkbare definitie, namelijk "toegevoegde waarde is een indicator voor de relatieve prestaties (of ontwikkeling) van leerlingen in een school in vergelijking tot andere scholen in dezelfde steekproef, waarbij een correctie plaatsvindt voor verschillen tussen leerlingen bij binnenkomst van de school die een bijdrage kunnen leveren aan de ontwikkeling van leerlingen." Sinds de ontwikkeling van statistische modellen die het schatten van toegevoegde waarde mogelijk hebben gemaakt, stapelt de literatuur waarin toegevoegde waarde wordt genoemd als een redelijke maat voor de schatting van effecten van scholen op de leerprestaties van leerlingen zich op.

De belangrijkste doelen van het onderzoek in dit proefschrift zijn a) het ontwikkelen van toegevoegde waarde indicatoren die gebruikt kunnen worden binnen de context van het Nederlandse onderwijstoezicht, b) het onderzoeken van de mate waarin toegevoegde waarde gebruikt kan worden bij het maken van een valide en eerlijke vergelijking van de effectiviteit van scholen en c) het onderzoeken van de mate

waarin toegevoegde waarde gebruikt kan worden in risico gestuurd onderwijstoezicht bij het voorspellen van prestaties van scholen in de toekomst. In elk van de studies in het proefschrift wordt een aspect van toegevoegde waarde in meer detail onderzocht, binnen diverse onderwijssectoren, waarbij aandacht is voor de bruikbaarheid van toegevoegde waarde binnen onderwijstoezicht en sterke en zwakke punten van de indicator.

Samenvatting van de belangrijkste resultaten van de empirische studies

In hoofdstuk 2 worden zowel de conceptuele als de empirische verschillen besproken tussen toegevoegde waarde-maten die in verschillende landen al gebruikt worden binnen onderwijstoezicht. Hoewel toegevoegde waarde in verschillende landen al wordt gebruikt als indicator voor opbrengsten binnen onderwijstoezicht, zijn er vele verschillende manieren waarop de toegevoegde waarde indicatoren binnen deze toezichtsystemen worden geoperationaliseerd. De verschillen in operationalisering komen met name tot uiting in de keuze van het statistische model (groeimodel of statusmodel) en de keuze van de controlevariabelen die worden gebruikt voor het maken van een eerlijke vergelijking tussen scholen. De keuze van de controlevariabelen bepaalt de conceptuele betekenis van de toegevoegde waarde indicator en het type school effect dat wordt geschat (bijvoorbeeld type A, B of type X school effect). De toegevoegde waarde-indicatoren die op dit moment binnen onderwijstoezicht worden gebruikt verschillen zeer sterk in de keuze voor controlevariabelen. In het Tennessee Value Added Assessment System (TVAAS), een indicator die in verschillende staten van de Verenigde Staten wordt toegepast, wordt alleen gecontroleerd voor verschillen tussen leerlingen in beginniveau. Bij Contextual Value Added (CVA), dat wordt gebruikt door Ofsted in Engeland, wordt daarentegen een veel uitgebreidere set van controlevariabelen gehanteerd. In deze laatste indicator worden het beginniveau, achtergrondkenmerken van de leerlingen en kenmerken van de schoolcompositie opgenomen als controlevariabelen. Een vergelijking van toegevoegde-waarde indicatoren in het voortgezet onderwijs die verschillen in de meegenomen controle variabelen op basis van het VOCL'99 cohort, laat zien dat er een sterke samenhang is tussen de geschatte toegevoegde waarde-indicatoren. Dit impliceert dat de indicatoren een vergelijkbaar achterliggend construct lijken te meten, ondanks de verschillen in controlevariabelen. Wanneer de toegevoegde waarde-indicatoren echter worden vergeleken op basis van de classificaties van scholen als ineffectief, gemiddeld en effectief worden de verschillen meer prominent. De classificatie van scholen in deze groepen en eventuele misclassificaties kunnen

belangrijke consequenties hebben voor individuele scholen, zoals meer intensief toezicht.

De effecten van leerling-mobiliteit en lange-termijn effecten van basisscholen zijn onderzocht in de studie gerapporteerd in hoofdstuk 3. Leerling-mobiliteit is het wisselen van scholen door leerlingen op een ander moment dan het begin en het einde van een opleiding. In traditionele toegevoegde waarde-indicatoren wordt een leerling aan de school voor voortgezet onderwijs toegekend waar hij of zij ook eindexamen heeft gedaan. In dergelijke indicatoren wordt er geen rekening mee gehouden dat een leerling op meerdere scholen voor voortgezet onderwijs kan hebben gezeten en dat de leerprestaties niet geheel aan de laatste school kunnen worden toegeschreven. Tevens wordt er geen rekening mee gehouden dat de basisschool waar een leerling op heeft gezeten een lange-termijn effect kan hebben, waar hij of zij voordeel van kan hebben tijdens het voortgezet onderwijs. Zowel leerling-mobiliteit als lange-termijn effecten van basisscholen zijn voorbeelden van imperfecte hiërarchische data. Het negeren van beide effecten kan mogelijk leiden tot vertekeningen van de schatting van de toegevoegde waarde van de school voor voortgezet onderwijs. Uit een empirische studie op basis van het VOCL'99 cohort blijkt dat er vertekeningen ontstaan in de schatting van de toegevoegde waarde van middelbare scholen door het negeren van leerling-mobiliteit. Hoewel de correlaties tussen een traditionele toegevoegde waarde-indicator en toegevoegde waarde waarbij rekening is gehouden met leerling-mobiliteit wederom hoog zijn, vinden er met name voor de scholen rondom het gemiddelde veel verschuivingen plaats. Voor meer dan 50% van de scholen worden er verschuivingen van meer dan 10 plaatsen in de rangorde gevonden. De geschatte toegevoegde waarde voor de meest en minst effectieve scholen lijkt vrij stabiel over de beide toegevoegde waarde-maten. In deze steekproef hebben we geen effecten gevonden van lange termijn effecten van de basisscholen op de schattingen van de toegevoegde waarde van scholen voor voortgezet onderwijs.

Een veel gehoorde kritiek op onderwijstoezicht en onderzoek naar onderwijseffectiviteit is de nadruk op leerprestaties. In hoofdstuk 4 wordt een empirische studie beschreven waarin de toegevoegde waarde van scholen voor voortgezet onderwijs wordt geschat voor zowel het cognitieve als niet-cognitieve domein. Toetsresultaten van leerlingen in het derde jaar van het voortgezet onderwijs (VOCL'99) zijn gebruikt om de toegevoegde waarde van scholen in kaart te brengen voor Nederlands, wiskunde, prestatiemotivatie en de door de leerling waargenomen sfeer in de klas. Deze studie bevestigt het beeld dat er meer verschillen zijn tussen scholen in de toegevoegde waarde voor Nederlands en wiskunde dan voor de niet-cognitieve uitkomstmaten. Met name de toegevoegde waarde van scholen op

prestatiemotivatie lijkt klein. Wanneer de toegevoegde waarde-indicatoren van de verschillende uitkomstmaten met elkaar worden vergeleken valt op dat er sprake is van matige positieve samenhangen tussen de toegevoegde waarde van Nederlands en wiskunde en tussen de toegevoegde waarde van Nederlands en sfeer in de klas. Dit betekent dat een school die goed presteert voor Nederlands vaak ook goed presteert voor wiskunde en voor de sfeer in de klas. De associatie tussen de toegevoegde waarde van prestatiemotivatie en de overige uitkomstmaten zijn over het algemeen klein en positief. Deze resultaten impliceren tevens dat effectiviteit van scholen binnen het cognitieve domein niet noodzakelijkerwijs ten koste gaat van de prestaties van scholen in het niet-cognitieve domein. De lage tot matige positieve correlaties hebben belangrijke consequenties voor het gebruik van toegevoegde waarde binnen het onderwijstoezicht. Dit betekent dat men niet uit kan gaan van één uitkomstmaat, maar dat er gebruik gemaakt zal moeten worden van verschillende uitkomstmaten om een volledig beeld te krijgen van de effectiviteit van een school. Dit combineren van meerdere indicatoren kent echter wel andere problemen, aangezien het bepalen van een valide of eerlijke grens of een school wel of niet voldoende presteert complexer wordt naarmate er sprake is van meer indicatoren.

Toegevoegde waarde indicatoren die gebruikt worden binnen het onderwijstoezicht worden vooral geschat op basis van toets- en examenresultaten van leerlingen. Een exploratie naar het schatten van de toegevoegde waarde van scholen voor voortgezet onderwijs op basis van de onderwijspositie wordt beschreven in het vijfde hoofdstuk. Het gebruik van toetsresultaten voor de beoordeling van scholen binnen onderwijstoezicht kan leiden tot een aantal perverse prikkels voor scholen om de scores op deze toetsen kunstmatig te verhogen. Voorbeelden van strategieën die scholen hiervoor kunnen gebruiken zijn voorzeggen tijdens de testafname, het laten gebruiken van extra hulpmiddelen en de zwakste leerlingen uitsluiten van toets-deelname. Het gebruiken van de onderwijspositie als uitkomstmaat in toegevoegde waarde heeft een aantal voordelen ten opzichte van het gebruik van toets-scores. In de eerste plaats heeft de onderwijspositie of een behaald diploma een grote maatschappelijke waarde. In de tweede plaats is het mogelijk om de prestaties van een school in één indicator uit te drukken in een gedifferentieerd onderwijssysteem, omdat iedere leerling op basis van het schooltype en leerjaar een score toegekend kan krijgen die de positie in het onderwijssysteem weergeeft. Dit in tegenstelling tot toets-scores die vaak niet vergelijkbaar zijn voor de verschillende onderwijstypen. Tenslotte lijkt toegevoegde waarde op basis van onderwijsposities meer robuust tegen bepaalde vormen van strategisch gedrag van scholen of leidt zelfs tot tegengestelde prikkels. Daarom wordt gesteld dat een dergelijke indicator een waardevolle aanvulling kan zijn

op toegevoegde waarde indicatoren op basis van toetsresultaten. Uit een empirische analyse van VOCL'99 data blijkt dat 7,1% van de verschillen in onderwijsposities tussen leerlingen kan worden toegeschreven aan de bezochte middelbare scholen, nadat gecontroleerd is voor beginniveau, achtergrondkenmerken van leerlingen en schoolcompositie. Wanneer deze verschillen tussen scholen worden uitgedrukt in bijvoorbeeld de kans op het behalen van een diploma theoretische leerweg of hoger komen de verschillen tussen scholen prominenter naar voren. Zo heeft een gemiddelde leerling op een effectieve school een 35% grotere kans op het behalen van een diploma theoretische leerweg of hoger dan een vergelijkbare, gemiddelde leerling op een ineffectieve school. Tevens werden er differentiële schooleffecten gevonden voor socio-economische status en beginniveau. Dit betekent dat scholen verschillen in effectiviteit voor verschillende subgroepen van leerlingen. Deze differentiële schooleffecten zijn gerelateerd aan de structuur van de school in termen van schooltypen die worden aangeboden. Scholen die alleen algemeen onderwijs aanbieden (HAVO en VWO) laten sterke verbanden zien tussen SES en beginniveau enerzijds en het bereikte eindniveau anderzijds, terwijl scholen die alleen VMBO aanbieden zwakkere verbanden laten zien. Voor brede scholengemeenschappen is geen duidelijk patroon zichtbaar met betrekking tot de differentiële schooleffecten. Deze laatste resultaten laten zien dat een leerling met een lager beginniveau en/of ouders met een lager opleidingsniveau gebaat lijken bij een opleiding aan een categorale VMBO instelling.

De studie beschreven in hoofdstuk 6 exploreert mogelijkheden voor het schatten van toegevoegde waarde in het middelbaar beroepsonderwijs. Hoewel toegevoegde waarde-indicatoren sinds de jaren tachtig van de vorige eeuw ontwikkeld zijn voor het primair en voortgezet onderwijs, is er maar een zeer beperkt aantal studies dat de effectiviteit van opleidingen en instellingen in het beroepsonderwijs in kaart brengt. Een mogelijke verklaring hiervoor is de complexiteit van niveaus en opleidingen. Bestaande studies gebruiken veelal logistische of probit modellen waarbij de multilevel-structuur binnen onderwijs (deelnemers binnen opleidingen binnen instellingen) genegeerd wordt. Onderwijsnummerdata van bijna 90,000 deelnemers die in 2008 een instelling verlaten vanuit een beroepsopleidende leerweg (BOL) zijn gebruikt voor het schatten van de toegevoegde waarde van clusters van opleidingen en instellingen. Het hoogst behaalde diploma van deelnemers die een instelling verlaten hebben is als uitkomstmaat gebruikt. Multilevel analyse laat behoorlijke verschillen zien tussen clusters van opleidingen, maar bijna geen verschillen tussen onderwijsinstellingen. Dit impliceert dat er binnen de grote instellingen wel verschillen in effectiviteit bestaan, maar tussen de instellingen bijna niet. Ongeveer 14% van de

variantie in behaalde diploma's kan worden toegeschreven aan de clusters van opleidingen, nadat een correctie heeft plaatsgevonden voor het beginniveau en achtergrondkenmerken van de deelnemers. Wanneer clusters van opleidingen geclassificeerd worden als ineffectief, gemiddeld en effectief blijkt dat ongeveer 30% van de clusters effectief is en 20% van de clusters ineffectief is. De kans op het behalen van een diploma verschilt aanzienlijk tussen de effectieve en ineffectieve clusters van opleidingen. Tenslotte wordt in dit hoofdstuk nog een vergelijking gemaakt tussen een model zonder (bruto schooleffect) en een model met controle variabelen (netto schooleffect of toegevoegde waarde). Uit deze vergelijking blijkt dat 20% van de clusters van opleidingen een andere classificatie krijgt wanneer er naar de toegevoegde waarde wordt gekeken.

De laatste empirische studie van het proefschrift betreft een risicoanalyse in het primair onderwijs op basis van toegevoegde waarde (hoofdstuk 7). Een recente ontwikkeling binnen het onderwijstoezicht is een risico-gestuurde aanpak. Binnen een risico-gestuurde aanpak kan de intensiteit en de frequentie van het toezicht per school verschillen afhankelijk van het risico dat een school loopt op onderpresteren. Deze manier van werken wordt gezien als efficiënter aangezien een toezichthouder zich kan richten op de onderwijsinstellingen die het meeste risico lopen. In een risico analyse wordt informatie van scholen uit het verleden *(t-1, t-2, etc.)* gebruikt om de huidige prestaties van een school te voorspellen *(t)*. Vervolgens worden de resultaten van een dergelijk model gebruikt om de prestaties van scholen in de toekomst te voorspellen *(t+1)*. Informatie die gebruikt kan worden om de prestaties van een school te voorspellen zijn kenmerken van de studentpopulatie, docentpopulatie, docentmobiliteit, financiële situatie van de school, resultaten van inspectiebezoeken en voorgaande prestaties. In deze studie zijn twee risico's gedefinieerd, namelijk het onvoldoende presteren op eindniveau (bruto schooleffecten) en onvoldoende toegevoegde waarde (netto schooleffecten). Onderpresteren van scholen is vervolgens geoperationaliseerd als onvoldoende prestaties op één of beide risico's. Leerlingvolgsysteem data voor begrijpend lezen van 500 basisscholen is gebruikt voor het schatten van de prestaties van basisscholen voor drie opeenvolgende cohorten. Twee verschillende statistische methoden zijn vervolgens gebruikt om onderpresteren van basisscholen te voorspellen, namelijk discriminant analyse en *regression tree* analyse. Beide methoden laten zien dat een combinatie van kenmerken van de leerlingpopulatie, voorgaande prestaties van de school en resultaten van inspectiebezoeken leiden tot de beste voorspelling van onderpresteren. Beide methoden laten tevens zien dat het noodzakelijk is om een zeer grote groep scholen verder te onderzoeken om bijna alle onderpresterende scholen te vinden. Er wordt

echter ook een groep scholen (40%) gevonden waarbij sprake is van een zeer klein risico op onderpresteren. Uit een check van de robuustheid van de risicomodellen door het toepassen op een later cohort blijkt dat wederom een grote groep onderpresterende scholen gevonden kan worden.

Algemene conclusies en beperkingen van het onderzoek

Tenslotte wordt in het laatste hoofdstuk van het proefschrift een samenvatting gegeven van de resultaten, alsmede algemene conclusies die over alle onderzoeken getrokken kunnen worden en beperkingen van het uitgevoerde onderzoek. In het onderzoek zijn drie typen data gebruikt (cohort data, onderwijsnummer-data en data uit leerlingvolgsystemen). Op basis van deze drie typen data is het mogelijk om de toegevoegde waarde van scholen te schatten. Echter, hiervoor moet data beschikbaar zijn op het niveau van de leerling en moet er minimaal een indicatie van het begin- en eindniveau aanwezig zijn. De drie typen data die in de verschillende studies zijn gebruikt hebben elk voor- en nadelen en leiden soms tot verschillende interpretaties van de toegevoegde waarde indicator. Zowel de cohortdata voor het voortgezet onderwijs als de onderwijsnummer-data in het beroepsonderwijs zijn uitermate geschikt voor het schatten van toegevoegde waarde op basis van zogenaamde statusmodellen. Dit zijn modellen waarbij het eindniveau van de leerlingen wordt gecorrigeerd voor het beginniveau en andere kenmerken van de leerlingen. Beide datasets beschikken over verschillende achtergrondkenmerken van leerlingen. Toegevoegde waarde op basis van een dergelijk model kan worden geïnterpreteerd als een indicator voor de relatieve *prestaties* van leerlingen in een school in vergelijking tot andere scholen in dezelfde steekproef. De data afkomstig uit leerlingvolgsystemen is meer geschikt voor het schatten van de toegevoegde waarde op basis van groeimodellen. In deze data zijn achtergrondgegevens van leerlingen bijna niet beschikbaar, maar daarentegen zijn er op verschillende momenten indicaties van de prestaties van leerlingen gemeten op eenzelfde latente schaal. Dit biedt de mogelijkheid om de groei van leerlingen in kaart te brengen. Toegevoegde waarde op basis van een groeimodel kan daardoor geïnterpreteerd worden als een indicator voor de relatieve *groei/ontwikkeling* van leerlingen in een school in vergelijking tot andere scholen in dezelfde steekproef.

Een tweede algemene bevinding is dat de schatting van de toegevoegde waarde van een school samengaat met een relatieve grote onbetrouwbaarheid. Door deze grote onbetrouwbaarheid kun je drie groepen scholen van elkaar onderscheiden, namelijk ineffectieve scholen, gemiddelde scholen en effectieve scholen. Deze

onbetrouwbaarheid is niet uniek voor toegevoegde waarde, maar geldt voor alle prestatie-indicatoren van scholen. Een mogelijke manier om hiermee om te gaan is het combineren van twee of meer opeenvolgende cohorten. Op deze manier wordt meer informatie gebruikt bij het schatten van de toegevoegde waarde van een school. Deze manier is toegepast in hoofdstuk 7 van dit proefschrift.

Met de verschillende studies is getracht een indicatie te krijgen van de validiteit van toegevoegde waarde indicatoren door steeds een aspect hiervan in detail te onderzoeken. Voor het gebruik van dergelijke indicatoren in de context van het onderwijstoezicht is de validiteit één van de belangrijkste voorwaarden. Echter, uit een vergelijking van toegevoegde waarde indicatoren op basis van verschillende onderliggende statistische modellen kan niet direct worden geconcludeerd welke indicator het meest valide is, of de beste benadering biedt van het te schatten schooleffect. Daarbij leidt het vergelijken van toegevoegde waarde indicatoren tot een andere vraag, namelijk: "Waar legt men de grens op basis waarvan men kan concluderen dat twee of meer toegevoegde waarde indicatoren vergelijkbaar of verschillend zijn?" Verschillen en overeenkomsten tussen indicatoren kunnen zichtbaar gemaakt worden aan de hand van percentages of aantallen misclassificaties, verschuivingen in rangordening en correlaties. Een eenduidig antwoord kan niet zomaar worden gegeven hoeveel misclassificaties getolereerd kunnen worden. Dit is in het bijzonder belangrijk binnen de context van het onderwijstoezicht, waar misclassificaties voor individuele scholen belangrijke consequenties kunnen hebben, zoals intensiever toezicht.

In de meeste studies hebben we nog geen gebruik kunnen maken van datasets die ook bij onderwijstoezicht worden gebruikt, vanwege het ontbreken van het beginniveau of omdat de data niet op het niveau van de leerlingen beschikbaar was. De uitzondering is de studie in het beroepsonderwijs. Hierdoor hebben we moeten zoeken naar alternatieve databronnen, waarin kenmerken van leerlingen soms anders zijn gemeten of waarin andere variabelen beschikbaar zijn. Hierdoor zijn de resultaten van deze studies niet zondermeer te generaliseren naar toegevoegde waarde indicatoren die in de toekomst binnen het onderwijstoezicht worden ontwikkeld.

Een terugkerend fenomeen in de verschillende studies was het voorkomen van missende waarden op verschillende controlevariabelen. Missende waarden voor controlevariabelen zijn een belangrijke beperking voor de validiteit van toegevoegde waarde, omdat het vaak de zwakkere leerlingen zijn waarvan data ontbreekt (Rubin et al., 2004). Het schatten van toegevoegde waarde op basis van leerlingen met volledige records resulteert daardoor in een overschatting van de prestaties van de leerlingen en

de school (Thomas et al., 1997b). In het derde hoofdstuk van dit proefschrift is gebruik gemaakt van een multilevel imputatiemethode, waardoor het probleem van missende waarden deels opgelost kan worden. Echter, dergelijke methoden zijn gebaseerd op de onderliggende assumptie, dat het ontbreken van waarde random is. In de nationale datasets, voor voortgezet en beroepsonderwijs, zijn de achtergrondkenmerken voor (bijna) alle leerlingen opgeslagen, maar de beschikbaarheid van het beginniveau kan nog een probleem zijn. Een voorbeeld hiervan zijn studenten die een opleiding in het beroepsonderwijs gaan volgen vanuit het beroepsleven. Voor deze studenten zal het achterhalen van een accuraat beginniveau niet altijd mogelijk zijn.

De focus van de studies in dit proefschrift ligt op bruikbaarheid van toegevoegde waarde binnen de context van het onderwijstoezicht. Alternatieve toepassingen van toegevoegde waarde zijn schoolkeuze, school verbetering en horizontale verantwoording. Elk van deze toepassingen van toegevoegde waarde stelt andere voorwaarden aan de indicator en het gebruik ervan. Door verschillen in voorwaarden kunnen de resultaten van de studies in dit proefschrift niet zomaar gegeneraliseerd worden naar overige toepassingen van toegevoegde waarde.

Aanbevelingen voor onderwijstoezicht

Een aantal aanbevelingen voor het gebruik van toegevoegde waarde kan worden gedaan voor het gebruik van toegevoegde waarde binnen het onderwijstoezicht. In de eerste plaats zal een weloverwogen en breed gedragen keuze gemaakt moeten worden met betrekking tot de controlevariabelen die opgenomen moeten worden in een toegevoegde waarde indicator. De conceptuele betekenis van de verschillende toegevoegde waarde modellen, verschillen in classificaties en de implicaties van het meenemen van controle variabelen voor groepen leerlingen kunnen een rol spelen in de te maken keuze.

De studie in het derde hoofdstuk laat zien dat het negeren van leerling-mobiliteit leidt tot vertekeningen in de schatting van de toegevoegde waarde van scholen in het voortgezet onderwijs. Echter, dit speelt voornamelijk voor scholen rondom het gemiddelde. Bij de resultaten van de minst effectieve scholen, die in het bijzonder interessant zijn in de context van het onderwijstoezicht, waren er geen verschuivingen zichtbaar. Op basis van deze bevinding zou men ervoor kunnen kiezen om leerling-mobiliteit niet te modelleren, aangezien het leidt tot een zeer complex en weinig transparant achterliggend statistisch model.

Ten derde blijkt dat scholen minder lijken te verschillen in hun toegevoegde waarde voor het niet-cognitieve domein. Daarnaast zijn in het vierde hoofdstuk van dit proefschrift matige positieve samenhangen gevonden tussen de toegevoegde waarde van scholen binnen het cognitieve domein en de toegevoegde waarde voor sfeer in de klas. De uitkomstmaten in deze studie zijn geen definitieve set en er zal onderzoek gedaan moeten worden naar verschillende andere mogelijke uitkomstmaten die relevant zijn voor het schatten van toegevoegde waarde. Op basis van dit en het toekomstig onderzoek zal een keuze moeten worden gemaakt in hoeverre indicatoren op basis van niet-cognitieve uitkomstmaten worden opgenomen binnen het onderwijstoezicht. De volgende punten kunnen in de overweging worden meegenomen: 1) de omvang van verschillen tussen scholen in de prestaties met betrekking tot de eventueel nieuwe uitkomstmaat, 2) het gebruik van meerdere uitkomstmaten leidt tot een meer gedetailleerd beeld van de prestaties van een school, en 3) het gebruik van meerdere uitkomstmaten leidt tot een meer complexe bepaling welke scholen goed presteren of niet.

In de studie in hoofdstuk 5 is een toegevoegde waarde model verkend op basis van de onderwijsposities van leerlingen. Uit de resultaten blijkt dat er aanzienlijke verschillen zijn tussen scholen voor voortgezet onderwijs in hun toegevoegde waarde voor onderwijsposities. Tevens blijkt dat er sprake is van differentiële schooleffecten voor beginniveau en socio-economische status. Een implicatie van deze resultaten is dat het niet mogelijk is om één indicator te gebruiken die het effect van de school op de onderwijspositie van de leerlingen accuraat weergeeft. Indicatoren voor meerdere subgroepen van leerlingen zijn nodig om een accuraat en gedetailleerd beeld te krijgen van de effectiviteit van een school. Toegevoegde waarde op basis van onderwijsposities lijkt meer robuust tegen bepaalde vormen van strategisch gedrag en daarom kan men overwegen om het te gebruiken naast indicatoren op basis van test- of examenresultaten. Het gebruik van meerdere indicatoren komt de robuustheid van onderwijstoezicht over het algemeen ten goede aangezien de verschillende indicatoren vaak leiden tot tegengestelde belangen voor scholen (Koretz, 2003).

Verder blijkt uit een studie naar toegevoegde waarde in het beroepsonderwijs dat verschillen in prestaties tussen deelnemers meer gerelateerd zijn aan de verschillende clusters van opleidingen dan aan de grote onderwijsinstellingen. Dit impliceert dat onderwijsinstellingen slechts zeer beperkt van elkaar verschillen in de mate waarin deelnemers diploma's behalen, maar dat binnen de instellingen verschillen zichtbaar zijn in de effectiviteit van clusters van opleidingen. In het huidige toezichtsysteem voor het beroepsonderwijs worden opbrengsten bepaald op het niveau van de onderwijsinstellingen. Op basis van de resultaten van het onderzoek in dit proefschrift

kan men zich afvragen in hoeverre het niveau van de onderwijsinstellingen het juiste is.

Tenslotte, blijkt uit de risicoanalyse (hoofdstuk 7) dat een grote groep scholen in aanmerking komt voor verder onderzoek om bijna alle onderpresterende scholen te vinden. Dit impliceert dat een risicomodel, zoals beproefd in deze dissertatie, onderpresteren van scholen in de toekomst niet erg accuraat kan schatten. Echter er kan een groep scholen (40%) worden gevonden die een zeer kleine kans hebben op onderpresteren geven hun voorgaande prestaties en andere kenmerken. Dit heeft belangrijke consequenties voor het gebruik van een risico-gestuurde manier van werken binnen het onderwijstoezicht. Hoewel een risico model niet zo accuraat is als men mag hopen leidt het toepassen in onderwijs toezicht toch tot een verbetering van de efficiëntie.

References

Ackerman, B. P., Brown, E. D., and Izard, C. E. (2004). The relations between contextual risk, earned income, and the school adjustment of children from economically disadvantaged families. *Developmental Psychology, 40,* 204-216.

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.

Aitkin, M. and Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Royal Statistical Society, 149,* 1-43.

Amrein-Beardsley, A. (2008). Methodological concerns about education value-added assessment system. *Educational researcher, 37,* 65-75.

Amrein-Beardsley, A., Berliner, D. C., and Rideau, S. (2010). Cheating in the first, second and third degree; Educators' responses to high stakes testing. *Education Policy Analysis Archives, 18,* 1-35.

Armstrong, D. and McVicar, D. (2000). Value added in further education and vocational training in Northern Ireland. *Applied Economics, 32,* 1727-1736.

Atkinson, J. W. & Reitman, W. (1958). Performance as a function of motive strength and expectancy of goal attainment. In J.W.Atkinson (Ed.), *Motives in fantasy, action and society* (pp. 278-287). Princton: Van Nostrand.

Betebenner, D. W. (2007). *Estimation of student growth percentiles for the Colorado Student Assessment Program* Dover, New Hampshire: National Centre for the Improvement of Educational Assessment (NCIEA).

Betebenner, D. W. (2009). *Growth, standards and accountability* Denver: Colorado Department of Education: The centre for assessment.

Bosker, R. J., Béguin, A., & Rekers-Mombarg, L. (2001). Hoe meten we de prestatie van een school? In A.B.Dijkstra, S. Karsten, R. Veenstra, & A. J. Visscher (Eds.), *Het oog der natie: Scholen op rapport* (pp. 121-135). Assen: Koninklijke Van Gorcum BV.

Bosker, R. J., Lam, J. F., Dekkers, H., & Vierke, H. (1997). *De betekenis van kwaliteits verschillen tussen basisscholen*. Enschede: Universiteit Twente.

Bosker, R. J., Lam, J. F., Luyten, H., Steen, R., & Vos, H. d. (1998). *Het vergelijken van scholen*. Enschede: Universiteit Twente.

Bosker, R. J. and Luyten, H. (2000). De stabiliteit en consistentie van differentiële schooleffecten. *Tijdschrift voor onderwijsresearch, 24,* 308-321.

Bosker, R. J. & Van der Velden, R. K. W. (1985). *Schooleffecten en rendementen* Groningen: RION: Instituut voor Onderwijsonderzoek Rijksuniversiteit Groningen.

Bosker, R. J. & Van der Velden, R. K. W. (1989). *The effects of secondary schools on the educational careers of disadvantaged pupils*. Groningen: RION, Institute for Educational Research.

Bradley, R. H. and Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology, 53,* 371-399.

Bressoux, P. and Bianco, M. (2004). Long-term teacher effects on pupils' learning gains. *Oxford Review of Education, 30,* 327-345.

Browne, W. J. (2009). *MCMC estimation in MLwiN Version 2.13*. Bristol: Centre for Multilevel Modelling, University of Bristol.

Browne, W. J. and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis, 1,* 473-514.

Browne, W. J., Goldstein, H., and Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling, 1,* 103-124.

Campbell, D. T. (1976). *Assessing the impact of planned social change* Kalamazoo, Michigan: Western Michigan University, Evaluation Center.

Cantrell, S., Fullerton, J., Kane, T. J., & Staiger, D. O. (2007). *National Board Certification and teacher effectiveness: Evidence from a random assignment experiment*.

Cervini, R. (2005). The relationship between school composition, school process and mathematics achievement in secondary education in Argentina. *International Review of Education, 51,* 173-200.

Coates, H. (2009a). Building quality foundations: indicators and instruments to measure the quality of vocational education and training. *Journal of Vocational Education and Training, 61,* 517-534.

Coates, H. (2009b). What's the difference? A model for measuring the value added by higher education in Australia. *Higher Education Management and Policy, 21,* 77-95.

Coe, R. and Fitz-Gibbon, C. T. (1998). School effectiveness research: Criticisms and recommendations. *Oxford Review of Education, 24,* 421-438.

Creemers, B., Kyriakides, L., & Sammons, P. (2010). *Methodological advances in educational effectiveness research: Quantitative methodology series* Routledge: New York.

Creemers, B. & Sleegers, P. (2003). De school als organisatie. In N.Verloop & J. Lowyck (Eds.), *Onderwijskunde* (pp. 112-122). Groningen: Wolters-Noordhoff.

Cullen, J. B. & Reback, R. (2006). *Tinkering toward accolades: School gaming under a performance accountability system* Cambridge: National Bureau of Economic Research.

De Fraine, B., Van Damme, J., Van Landeghem, G., Opdenakker, M-C., and Onghena, P. (2003). The effect of schools and classes on language achievement. *British Educational Research Journal, 29,* 841-859.

de Jong, P. (2012). The Health Impact of Mandatory Bicycle Helmet Laws. *Risk Analysis, 32,* 782-790.

de Nationale ombudsman (2009). *Onderwijsvrijheid en onderwijstoezicht in het radicaal vernieuwend onderwijs* Den Haag: de Nationale ombudsman.

De Wolf, I. and Verkroost, J. J. H. (2011). Evaluatie van de theorie en praktijk van het nieuwe onderwijstoezicht. *Tijdschrift voor Toezicht, 2,* 7-24.

Dekkers, H. P. J. M., Bosker, R. J., and Driessen, G. W. J. M. (2000). Complex inequalities of educational opportunities: A large scale longitudinal study on the relation between gender, social class, ethnicity and school succes. *Educational Research and Evaluation, 6,* 59-82.

Dijkstra, A. B., Karsten, S., Veenstra, R., & Visscher, A. J. (2001). *Het oog der natie: Scholen op rapport*. Assen: Koninklijke Van Gorcum BV.

Doolaard, S. & Leseman, P. P. M. (2008). *Versterking van het fundament: Integrerende studie n.a.v. de opbrengsten van de onderzoekslijn Sociale en Institutionele context van scholen uit het Onderzoeksprogramma beleidsgericht onderzoek primair onderwijs 2005-2008*. Groningen: GION.

Downey, D. B., Von Hippel, P. T., and Hughes, M. (2008). Are "failing" schools really failing? Removing the influence of non-school factors from measures of school quality. *Sociology of Education, 81,* 242-270.

Duncan, G. J. and Brooks-Gunn, J. (2000). Family poverty, welfare reform and child development. *Child Development, 71,* 188-196.

Ehren, M. C. M, De Leeuw, J., and Scheerens, J. (2005). On the Impact of the Dutch Educational Supervision Act. Analyzing Assumptions Concerning the Inspection of Primary Education. *American Journal of Evaluation, 26,* 60-76.

Elte, R. & Scholtes, E. (2002). *Uit de luwte, over strategische veranderingen in en rond de onderwijsinspectie 1990-2000*. Utrecht: Inspectie van het Onderwijs.

Engec, N. (2006). Relationship between mobility of students and student performance and behavior. *The Journal of Educational Research, 99,* 167-178.

Evers, A., Van Vliet-Mulder, J. C., & Groot, C. J. (2000). *Documentatie van tests en testresearch in Nederland [Documentation of tests and test research in The Netherlands]*. Assen: Van Gorcum.

Feenstra, H., Kamphuis, F., Kleintjes, F., & Krom, R. (2010). *Wetenschappelijke verantwoording Begrijpend lezen voor groep 3 tot en met 6* Arnhem: CITO.

Fielding, A. & Goldstein, H. (2006). *Crossclassified and multiple membership structures in multilevel models: An introduction and review* Nottingham: Department for education and skills.

Figlio, D. N. & Getzler, L. S. (2002). *Accountability, ability and disability: Gaming the system* Cambridge: National Bureau of Economic Research.

Freiberg, H. J. (1996). From tourists to citizens in the classroom. *Educational Leadership, 54,* 32-36.

Gibb, H. J. (1997). Epidemiology and cancer risk assessment. In V.Molak (Ed.), *Fundamentals of risk analysis and risk assessment* (Boca Raton, New York, London, Tokyo: Lewis Publishers.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement, 8,* 369-395.

Goldstein, H. (2001). Using pupil performance data for judging schools and teachers: scope and limitations. *Britisch Educational Research Journal, 27,* 433-442.

Goldstein, H., Burgess, S., and McConnell, B. (2007). Modelling the effect of pupil mobility on school differences in educational achievement. *Royal Statistical Society, 170,* 941-954.

Goldstein, H., Kounali, D., and Robinson, A. (2008). Modelling measurement errors and category misclassifications in multilevel models. *Statistical Modelling, 8,* 243-261.

Goldstein, H. and Sammons, P. (1997). The influence of secondary and junior schools on sixteen year examination performance: A cross-classified multilevel analysis. *School Effectiveness and School Improvement, 8,* 219-230.

Goldstein, H. and Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, Series A, 159,* 385-443.

Gorard, S. (2006). Value added is of little value. *Journal of Education Policy, 21,* 235-243.

Gray, J. (2004a). Frames of reference and traditions of interpretation: Some issues in the identification of "under-achieving" schools. *British Journal of Educational Studies, 52,* 293-309.

Gray, J. (2004b). School effectiveness and the 'other outcomes' of secondary schooling: a reassessment of three decades of British research. *Improving Schools, 7,* 185-198.

Gray, J., Peng, W. J., Steward, S., and Thomas, S. (2004). Towards a typology of gender-related school effects: some new perspectives on a familiar problem. *Oxford Review of Education, 30,* 529-550.

Gutman, L. M., Sameroff, A. J., and Cole, R. (2003). Academic growth curve trajectories from 1st grade to 12th grade: Effects of multiple social risk factors and preschool child factors. *Developmental Psychology, 39,* 777-790.

Haertel, G. D., Walberg, H. J., and Haertel, E. H. (1981). Socio-psychological environments and learning: a quantitative synthesis. *British Educational Research Journal, 7,* 27-36.

Harmon, C. M. (1992). Value added assessment. In D.D.Bragg (Ed.), *Alternative approaches to outcomes assessment for post secondary vocational education* (Berkeley: National Center for Reserach in Vocational Education.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London and New York: Routledge.

Hedeker, D. (2008). Multilevel Models for Ordinal and Nominal Variables. In J.De Leeuw & E. Meijer (Eds.), *Handbook of Multilevel Analysis* (pp. 237-274). New York: Springer.

Hill, P. W. and Goldstein, H. (1998). Multilevel modeling of educational data with cross-classification and missing identification for units. *Journal of Educational and Behavioral Statistics, 23,* 117-128.

Hill, P. W. and Rowe, K. J. (1996). Multilevel modelling in school effectiveness research. *School Effectiveness and School Improvement, 7,* 1-34.

Hill, P. W. and Rowe, K. J. (1998). Modelling student progress in studies of educational effectiveness. *School Effectiveness and School Improvement, 9,* 310-333.

Hofman, R. H., Hofman, W. H., and Guldemond, H. (1999). Social and cognitive outcomes: A comparison of contexts of learning. *School Effectiveness and School Improvement, 10,* 352-366.

Hox, J. J. (2002). *Multilevel analysis: Techniques and apllications*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Hox, J. J. & Roberts, J. K. (2011). *Handbook of advanced multilevel analysis*. New York, London: Routledge Taylor and Francis Group.

Hulett, D. T. & Preston, J. Y. (2000). Garbage In,Garbage Out? Collect Better Data for Your Risk Assessment. In *Proceedings of the Project Management Institute Annual Seminars & Symposium* (pp. 983-989). Houston.

Hustinx, P. W. J., Kuyper, H., Van der Werf, M. P. C., and Dijkstra, P. (2009). Achievement motivation revisited: new longitudinal data to demonstrate its predictive power. *Educational Psychology, 29,* 561-582.

Inspectie van het Onderwijs (2003). *Het bepalen van de toegevoegde waarde door basisscholen* Utrecht: Inspectie van het Onderwijs.

Inspectie van het Onderwijs (2006). *Wetenschappelijke onderbouwing van het waarderingskader PO 2005* Utrecht: Inspectie van het Onderwijs.

Inspectie van het Onderwijs (2009). *Toezichtkader PO/VO 2009* Utrecht: Inspectie van het Onderwijs.

Inspectie van het Onderwijs (2011a). *Addendum toezichtkader bve 2012: Beoordeling opbrengsten bekostigd en niet-bekostigd mbo-onderwijs per 1 januari 2012* Utrecht: Inspectie van het Onderwijs.

Inspectie van het Onderwijs (2011b). *Analyse en waarderingen van opbrengsten primair onderwijs* Utrecht: Inspectie van het Onderwijs.

Inspectie van het Onderwijs (2011c). *Opbrengsten overzicht 2011: Toelichting* Utrecht: Inspectie van het onderwijs.

Inspectie van het Onderwijs (2011d). *Opbrengstenkaart 2011: Technische toelichting* Utrecht: Inspectie van het Onderwijs.

Inspectie van het Onderwijs (2011e). *Selectief en slagvaardig, werken met de WOT (2000-2010)* Utrecht: Inspectie van het Onderwijs.

Inspectie van het Onderwijs (2011f). *Toezichtkader BVE 2012* Utrecht: Inspectie van het Onderwijs.

Inspectie van het Onderwijs (2012). *De staat van het onderwijs. Onderwijsverslag 2010/2011* Utrecht: Inspectie van het Onderwijs.

Inspectorate of Education (2009). *Risk-based inspection as of 2009*. Utrecht: Inspectorate of Education.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in Chicago public schools. *Journal of public economics, 89,* 761-796.

Jacob, B. A. and Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics, 118,* 843-877.

Jacob, M. & Tieben, N. (2007). *Social selectivity of track mobility in secondary schools. A comprison of intra-secondary transitions in Germany and the Netherlands* Mannheim: Mannheimer zentrum for Europaïsche sozialforschung.

Jones, B. D. (2008). The unintended outcomes of high-stakes testing. *Journal of applied school psychology, 23,* 65-86.

Kane, T. J. & Staiger, D. O. (2008). *Are teacher-level value added estimates biased? An experimental validation of non-experimental estimates.* NBPTS [preliminary draft].

Keeney, R. L. and von Winterfeldt, D. (2011). A Value Model for Evaluating Homeland Security Decisions. *Risk Analysis, 31,* 1470-1487.

Keeves, J. P., Hungi, N., and Afrassa, T. (2005). Measuring value added effect across schools: Should schools be compared in performance? *Studies in Educational Evaluation, 31,* 247-266.

Knuver, J. W. M. (1993). *De relatie tussen klas- en schoolkenmerken en het affectief functioneren van leerlingen [The relation between class- and school characteristics and affective functioning of students]*. Groningen: RION.

Konu, A. I., Lintonen, T. P., and Autio, V. I. (2002). Evaluation of Well-being in Schools - A multilevel analysis of general subjective well-being. *School Effectiveness and School Improvement, 13,* 187-200.

Koretz, D. M. (2003). Using multiple Measures to address perverse incentives and score inflation. *Educational Measurement: Issues and Practice, 22,* 18-26.

Koretz, D. M. (2005). *Alignment, high stakes and the inflation of test scores* Los Angeles: National Center for Research in Evaluation.

Koretz, D. M. (2008). A measured approach: Value-added models are a promising improvement, but no one measure can evaluate teacher performance. *American Educator, 32,* 18-39.

Kuyper, H. & Van der Werf, M. P. C. (2003a). *VOCL'99-1: Technisch rapport [Technical report]* Groningen: GION.

Kuyper, H. & Van der Werf, M. P. C. (2003b). *VOCL'99: de resultaten van het eerste leerjaar [VOCL'99: results of the first grade]* Groningen: GION.

Kuyper, H., Van der Werf, M. P. C., and Lubbers, M. J. (2010). Motivation, meta-cognition and self-regulation as predictors of long term educational attainment. *Educational Research and Evaluation, 6,* 181-205.

Kyriakides, L. (2004). Differential school effectiveness in relation to sex and social class: Some implications for policy evaluation. *Educational Research and Evaluation, 10,* 141-161.

Kyriakides, L. and Creemers, B. (2008). A longitudinal study on the stability over time of school and teacher effects on student outcomes. *Oxford Review of Education, 34,* 521-545.

Landis, J. R. and Koch, G. G. (1977). Measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

Leckie, G. (2008). *Modelling the effects of pupil mobility and neighbourhood on school differences in educational achievement* Bristol: CMPO, University of Bristol.

Leckie, G. (2009). The complexity of school and neighbourhood effects and movements of pupils on school differences in models of educational achievement. *Journal of the Royal Statistical Society, 172,* 537-554.

Leckie, G. and Goldstein, H. (2009). The limitations of using school lague tables to inform school choice. *Journal of the Royal Statistical Society, Series A, 172,* 835-851.

Leckie, G. and Goldstein, H. (2011a). A note on: The limitations of school league tables to inform school choice. *Journal of the Royal Statistical Society, Series A, 174,* 833-836.

Leckie, G. and Goldstein, H. (2011b). Understanding uncertainty in school league tables. *Fiscal Studies, 32,* 207-224.

Lee, K. & Weimer, D. (2002). *Building value added assessment into Michigan's Accountability system: Lessons from other states* Michigan: The Education Policy Center, Michigan State University.

Lemke, R. J., Hoerandner, C. M., and McMahon, R. E. (2006). Student assessment, non-test-takers, and school accountability. *Education Economics, 14,* 235-250.

Leventhal, T. and Brooks-Gunn, J. (2004). A randomized study of neighborhood effects on low-income children's educational outcomes. *Developmental Psychology, 40,* 488-507.

Lewis, C. C., Schaps, E., and Watson, M. S. (1996). The caring classroom's academic edge. *Educational Leadership, 54,* 16-21.

Luyten, H. (1994). Stability of school effects in secondary education. In *American Educational Research Association Annual Meeting in New Orleans* University of Twente, Department of Education.

Luyten, H. (1998). School effectiveness and student achievement consistent across subjects? Evidence from Dutch elementary and secondary education. *Educational Research and Evaluation, 4,* 281-306.

Luyten, H. (2003). The size of schools effects compared to teacher effects: An overview of the research literature. *School Effectiveness and School Improvement, 14,* 31-51.

Luyten, H., Visscher, A. J., and Witziers, B. (2005). School effectiveness research: From a review of the criticism to recommendations for further development. *School Effectiveness and School Improvement, 16,* 249-279.

Ma, X. (2001). Stability of school academic performance across subject areas. *Journal of Educational Measurement, 38,* 1-18.

Mandeville, G. K. (1988). School effectiveness indices revisited: Cross-year stability. *Journal of Educational Measurement, 25,* 349-356.

Mandeville, G. K. and Anderson, L. W. (1987). The stability of school effectiveness indices across grade levels and subject areas. *Journal of Educational Measurement, 24,* 203-216.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability* New York: RAND.

Mehana, M. and Reynolds, A. J. (2004). School mobility and achievement: a meta-analysis. *Children and Youth Services Review, 26,* 93-119.

Mertens, F. J. H. (2002). De wettelijke formulering van de inspectietaak: McKinsey en artikel 5 van de Wet op het Basisonderwijs1975-1982. *Nederlands Tijdschrift voor Onderwijsrecht en Onderwijsbeleid, 14,* 38-62.

Mertens, F. J. H. (2009). *De regulerende staat. Ontwikkelingen van het toezicht door Inspecties* Den Haag: Nederlandse School voor Openbaar Bestuur.

Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Educational Review, 16,* 283-301.

Ministry of Education, C. a. S. (2007). *The education system in the Netherlands* The Haque: Ministry of Education, Culture and Science.

Molak, V. (1997). *Fundamentals of risk analysis and risk management*. Boca Raton, New York, London, Tokyo: Lewis Publishers.

Mortimore, P. and Sammons, P. (1994). Schooleffectiveness and value added measures. *Assessment in Education: Principles, Policy & Practice, 1*.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School matters: The junior years*. Somerset: Open Books.

Nash, R. (2003). Is the school composition effect real? A discussion with evidence from the UK PISA data. *School Effectiveness and School Improvement, 14,* 441-457.

Neville, P. G. (1999). *Decision Trees for Predictive Modeling* SAS Institute.

Nichols, S. L. & Berliner, D. C. (2005). *The inevitable corruption of indicators and educators through high stakes testing* Tempe: Education olicy Research Unit (EPRU).

Nuttall, D., Goldstein, H., Prosser, R., and Rasbash, J. (1989). Differential school effectiveness. *International Journal of Educational Research, 13,* 769-776.

OECD (2008). *Measuring improvements in learning outcomes: Best practices to assess the value-added of schools*. OECD.

Ofsted (2010). *The evaluation schedule for schools*. Manchester: The Office of Standards in Education.

Ofsted (2011). *The framework for school inspection* Manchester: The Office for Standards in Education.

Onderwijsraad (1999). *Deugdelijk toezicht* Den Haag: Onderwijsraad.

Onderwijsraad (2002). *Vaste grond onder de voeten* Den Haag: Onderwijsraad.

Onderwijsraad (2003). *Wat scholen toevoegen.* Den Haag: Onderwijsraad.

Opdenakker, M-C. and Van Damme, J. (2000). Effects of schools, teaching staff and classes on achievement and well-being in secondary education: similarities and differences between school outcomes. *School Effectiveness and School Improvement, 11,* 165-196.

Opdenakker, M-C. and Van Damme, J. (2001). Relationship between school composition and characteristics of school process and their effect on mathematic achievement. *British Educational Research Journal, 27,* 407-432.

Opdenakker, M-C. and Van Damme, J. (2007). Do school context, student composition and school leadership affect school practice and outcomes in secondary education? *British Educational Research Journal, 33,* 179-206.

Peetsma, T., Van der Veen, I., Koopman, P., and Van Schooten, E. (2006). Class composition influences on pupils' cognitive development. *School Effectiveness and School Improvement, 17,* 275-302.

Peschar, J. L. (2004). Cross-curricular competencies: Developments in a new are of education outcome indicators. In J.H.Moskowitz & M. Stephens (Eds.), *Comparing learning outcomes* (pp. 79-107). NewYork: RoutledgeFalmer.

Peschar, J. L. & Van der Wal, M. (2001). Waarom alleen rapportcijfers of diploma's? In A.B.Dijkstra, S. Karsten, R. Veenstra, & A. J. Visscher (Eds.), *Het oog der natie: Scholen op rapport* (Assen: Koninklijke Van Gorcum BV.

Pustjens, H., Van de Gaer, E., Van Damme, J., Onghena, P., and Van Landeghem, G. (2007). The short-term and the long-term effect of primary schools and classes on mathematics and language scores. *British Educational Research Journal, 33,* 419-440.

Quené, H. and Van den Bergh, H. (2004). On multi-level modelling of data from repeated measures designs: a tutorial. *Speech Communication, 43,* 103-121.

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2009). *A user's guide to MLwiN.* Bristol: Centre for Multilevel Modelling, University of Bristol.

Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methological diversity. *Educational researcher, 34,* 25-31.

Raudenbush, S. W. and Bryk, A. S. (1986). A Hierarchical model for studying school effects. *Sociology of Education, 59,* 1-17.

Raudenbush, S. W. and Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20,* 307-335.

Ray, A. (2006). *School value added measures in England: A paper for the OECD project on the development of value-added models in education systems* Department of Education Skills.

Reynolds, D. & Teddlie, C. (1999a). The future agenda of studies into the effectiveness of schools. In R.J.Bosker, B. Creemers, & S. Stringfield (Eds.), *Enhancing educational excellence, equity and efficiency* (pp. 223-251). Dordrecht: Kluwer Academic Publishers.

Reynolds, D. & Teddlie, C. (1999b). The future agenda of studies into the effectiveness of schools. In R.J.Bosker, B. Creemers, & S. Stringfield (Eds.), *Enhancing educational excellence, equity, and efficiency: evidence from evaluations of systems and schools in change* (pp. 223-251). Dordrecht: Kluwer Academic Publishers.

Rodgers, T. (2005). Measuring value added in higher education? Do any of the recent experiences in secondary education in the United Kingdom suggest a way forward? *Quality Assurance in Education, 13,* 95-106.

Rodgers, T. (2007). Measuring value added in higher education: A proposed methodology for developing a performance indicator based on the economic value added to graduates. *Education Economics, 15,* 55-74.

Roede, E. (2001). Criteria voor schoolkwaliteit. In A.B.Dijkstra, S. Karsten, R. Veenstra, & A. J. Visscher (Eds.), *Het oog der natie: Scholen op rapport* (pp. 79-94). Assen: Koninklijke Van Gorcum BV.

Roeleveld, J. (2003a). *Herkomstkenmerken en begintoets:Secundaire analyses op het PRIMA-cohortonderzoek.*

Roeleveld, J. (2003b). *Herkomstkenmerken en begintoets:Secundaire analyses op het PRIMA-cohortonderzoek [Background characteristics and prior achievement: Secondary analisys on the PRIMA-cohort studies]* Amsterdam: SCO-Kohnstamm Instituut.

Rothstein, J. (2008). *Student sorting and bias in value added estimation: Selection on observables and unobservables* (Rep. No. CEPS workingpaper No.170).

Rubin, D. B., Stuart, E. A., and Zanutto, E. L. (2004). A potential outcomes view of value added assessment in education. *Journal of Educational and Behavioral Statistics, 29,* 103-116.

Salganik, L. H. (1994). Apples and apples: Comparing performance indicators for places with similar demographic characteristics. *Educational Evaluation and Policy Analysis, 16,* 125-141.

Sammons, P., Hillman, J., & Mortimore, P. (1995a). *Key characteristics of effective schools: A review of school effectiveness research* London: Office of Standards in Education.

Sammons, P., Nuttall, D., and Cuttance, P. (1993). Differential school effectiveness: Results from a reanalysis of the inner London Education Authority's junior school project data. *British Education Research Journal, 19,* 381-405.

Sammons, P., Nuttall, D., Cuttance, P., and Thomas, S. (1995b). Continuity of school effects: A longitudional analysis of primary and secondary school effects on GCSE performance. *School Effectiveness and School Improvement, 6,* 285-307.

Sammons, P., Thomas, S., & Mortimore, P. (1997). *Forging Links: Effective Schools and Effective Departments* London: Paul Chapman.

Sanders, W. L. (2003). Beyond No Child Left Behind. In *2003 Annual Meeting American Educational Research Association*.

Sanders, W. L. and Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8,* 299-311.

Saunders, L. (1999). A brief history of educational "Value added": How did we get to where we are? *School Effectiveness and School Improvement, 10,* 233-256.

Schagen, I. (2006). The use of standardized residuals to derive value-added measures of schoolperformance. *Educational Studies, 32,* 119-132.

Schagen, I. and Hutchison, D. (2003). Adding value in educational research - The marriage of data and analytical power. *British Educational Research Journal, 29,* 749-765.

Sharp, S. (2006). Assessing value-added in the first year of schooling: Some results and methodological considerations. *School Effectiveness and School Improvement, 17,* 329-346.

Snijders, T. & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: SAGE publications.

Snijders, T. A. B. & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Los Angeles / London / New Delhi / Singapore / Washington DC: SAGE Publications.

Solomon, D., Watson, M. S., Delucchi, K. L., Schaps, E., and Battistich, V. (1988). Enhancing children's prosocial behavior in the classroom. *American Educational Research Journal, 25,* 527-554.

Sparrow, M. K. (2000). *The regulatory craft; controling risks, solving problems and managing compliance*. Washington: Brookings Institution Press.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B, 64,* 583-639.

Steenbergen, H. (2009). *Vrije scholen en reguliere scholen vergeleken: Een onderzoek naar de effectiviteit van Vrije scholen en reguliere scholen voor voortgezet onderwijs*. Groningen: GION, Gronings Instituut voor Onderzoek van Onderwijs.

Strand, S. and Demie, F. (2007). Pupil mobility, attainment and progress in secondary school. *Educational Studies, 33,* 313-331.

Swanborn, M. & De Wolf, I. (2008). *Betrouwbaarheid van opbrengstmaten: Tussen- en eindresultaten in het basisonderwijs* Utrecht: Inspectie van het Onderwijs.

Tate, R. L. (2004). A cautionary note on shrinkage estimates of school and teacher effects. *Florida Journal of Educational Research, 42,* 1-21.

Teddlie, C. & Reynolds, D. (2000a). *The international handbook of school effectiveness research*. London: Falmer Press.

Teddlie, C., Reynolds, D., & Sammons, P. (2000b). The methodology and scientific proportions. In C.Teddlie & D. Reynolds (Eds.), *The international handbook of school effectiveness research* (pp. 55-133). London: Falmer Press.

Temple, J. A. and Reynolds, A. J. (1999). School mobility and achievement: longitudinal findings from an urban cohort. *Journal of School Psychology, 37,* 355-377.

Ten Dam, G. & Vermunt, J. (2003). De leerling. In N.Verloop & J. Lowyck (Eds.), *Onderwijskunde* (pp. 150-193). Groningen: Wolters-Noordhoff.

Teodorovic, J. (2011). Classroom and school factors related to student achievement: what works for students? *School Effectiveness and School Improvement, 22,* 215-236.

Thomas, S. (1998). Value-added measures of school effectiveness in the United Kingdom. *Prospects, 28,* 91-108.

Thomas, S. (2001). Dimensions of secondary school effectiveness: Comparative analyses across regions. *School Effectiveness and School Improvement, 12,* 285-322.

Thomas, S., Sammons, P., Mortimore, P., and Smees, R. (1997a). Differential secondary school effectiveness: Comparing the performance of different pupil groups. *British Educational Research Journal, 23,* 451-469.

Thomas, S., Sammons, P., Mortimore, P., and Smees, R. (1997b). Stability and consistency in secondary schools' effects on students' GCSE outcomes over three years. *School Effectiveness and School Improvement, 8,* 169-197.

Thomas, S., Smees, R., MacBeath, J., Robertson, P., and Boyd, B. (2000). Valuing pupils' views in Scottish schools. *Educational Research and Evaluation, 6,* 281-316.

Timmermans, A. C., Doolaard, S., and De Wolf, I. (2011). Conceptual and empirical differences among various value added models for accountability. *School Effectiveness and School Improvement, 22,* 393-413.

Timmermans, A. C., Snijders, T. A. B., & Bosker, R. J. (2012). *In search of value added in case of complex school effects*.

Tymms, P. (1995). The long-term impact of schooling. *Evaluation and Research in Education, 9,* 99-108.

Tymms, P. & Dean, C. (2004). *Value-added in the primary school league tables: A report for the National Association of Head Teachers* Durham: CEM Centre, University of Durham.

Tymms, P., Merrell, C., and Henderson, B. (2000). Baseline assessment and progress during the first three years at school. *Educational Research and Evaluation, 6,* 105-129.

Van Buuren, S. (2011). Multiple Imputation of Multilevel Data. In J.J.Hox & J. K. Roberts (Eds.), *Handbook of Advanced Multilevel Analysis* (pp. 173-196). New York: Routledge.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., and Rubin, D. B. (2006). Fully Conditional Specification of Multivariate Imputation. *Journal of Statistical Computation and Simulation, 76,* 1049-1064.

Van Damme, J., De Fraine, B., Van Landeghem, G., Opdenakker, M-C., and Onghena, P. (2002). A new study on educational effectiveness in secondary schools in Flanders: An introduction. *School Effectiveness and School Improvement, 13,* 383-397.

Van Damme, J., Opdenakker, M.-C., Van Landeghem, G., De Fraine, B., Pustjens, H., & Van de Gaer, E. (2006). *Educational effectiveness; An introduction to international and Flemish research on schools, teachers and classes* Leuven: Centre for Educational Effectiveness and Evaluation.

Van de Grift, W. (2009). Reliability and validity in measuring the value added of schools. *School Effectiveness and School Improvement, 20,* 269-285.

Van den Berghe, W. (1996). *Quality Issues and trends in vocational education and training in Europe* Thessaloniki: European Centre for the Development of Vocational Training.

Van den Berghe, W. (1997). *Indicators in perspective: The use of quality indicators in vocational education and training* Tessaloniki: European Centre for the Developement of Vocational Training.

Van der Werf, M. P. C. & Guldemond, H. (1996). *Omvang, stabiliteit en consistentie van schooleffecten in het basisonderwijs*. Groningen: GION.

Van Dijk, H. (1995). *Constructie en validering van de GIVO, Groninger Intelligentietest voor Voortgezet Onderwijs*. Lisse: Swets & Zeitlinger B.V.

Van Landeghem, G., Van Damme, J., Opdenakker, M-C., De Fraine, B., and Onghena, P. (2002). The effects of schools and classes on noncognitive outcomes. *School Effectiveness and School Improvement, 13,* 429-451.

Van Zolingen, S. J. (1995). *Gevraagd: Sleutelkwalificaties*. KUN, Nijmegen.

Van Zolingen, S. J. and KLaassen, C. A. (2003). Selection processes in a Delphi study about key qualifications in Senior Secondary Vocational Education. *Technological Forecasting and Social Change, 70,* 317-340.

Veenstra, R. (1999). *Leerlingen-klassen-scholen: Prestaties en vorderingen van leerlingen in het voortgezet onderwijs [Students-classes-schools: achievement and progress of students in secondary education]*. Groningen: ICS (Interuniversity Centre for Social Science Theory and Methodology).

Verhelst, N., Staphorsius, G., & Kleintjes, F. (2003). *Scholen langs de meetlat* Arnhem: CITO.

Veugelers, W. & De Kat, E. (1998). *Opvoeden in het voortgezet onderwijs. Leerlingen, ouders en docenten over de pedagogische opdracht en de afstemming tussen gezin en school.* Assen: Van Gorcum & Comp. BV.

Webster, W. J., Mendro, R. L., Orsak, T. H., & Weerasinghe, D. (1996). The applicability of selected regression and hierarchical linear models to the estimation of school and teacher effects. In *Annual meeting of the National Council on Measurement in Education*.

Webster, W. J., Mendro, R. L., Orsak, T. H., & Weerasinghe, D. (1998). An application of hierarchical linear modeling to the estimation of school and teacher effect. In *Annual meeting of the American Educational Research Association, april 13-17, 1998*.

Wijnstra, J., Ouwens, M., & Béguin, A. (2003). *De toegevoegde waarde van de basisschool* Arnhem: CITO.

Willms, J. D. (1986). Social class segregation and its relationship to pupils' examination results in Scotland. *American Sociological Review, 51,* 224-241.

Willms, J. D. (1992). *Monitoring school performance: A guide for educators*. Washington DC: The Falmer Press.

Willms, J. D. and Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement, 26,* 209-232.

Woodhouse, G., Yang, M., Goldstein, H., and Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society, 159,* 201-212.

Yen, S., Schafer, W. D., & Rahman, T. (1999). *School effect indices: Stability of one- and two-level formulations.*

Yunker, J. A. (2005). The dubious utility of the value-added concept in higher education: the case of accounting. *Economics of Education Review, 24,* 355-367.

Zijsling, D. H., Kuyper, H., Lubbers, M. J., & Van der Werf, M. P. C. (2005). *VOCL'99-3 Technisch rapport [VOCL'99-3 Technical Report]* Groningen: GION, Groningen Institute for educational research.

ICO Dissertation Series

In the ICO Dissertation Series the dissertations of graduate students from faculties and institutes on educational research within the ICO Partner Universities are published: Eindhoven University of Technology, Leiden University, Maastricht University, Open University of the Netherlands, University of Amsterdam, University of Twente, Utrecht University, VU University Amsterdam, and Wageningen University, and formerly University of Groningen (until 2006), Radboud University Nijmegen (until 2004), and Tilburg University (until 2002). The University of Groningen, University of Antwerp, University of Ghent, and the Erasmus University Rotterdam have been 'ICO 'Network partner' in 2010 and 2011. From 2012 onwards, these ICO Network partners are full ICO partners, and from that period their dissertations will be added to this dissertation series.

ICO Dissertations 2011/2010:

235. Van Stiphout, I.M. (14-12-2011). *The development of algebraic proficiency.* Eindhoven: Eindhoven University of Technology.

234. Elffers, L. (14-12-2011). *The transition to post-secondary vocational education: Students' entrance, experiences, and attainment.* Amsterdam: University of Amsterdam.

233. Cornelissen, L.J.F. (29-11-2011). *Knowledge processes in school-university research networks.* Eindhoven: Eindhoven University of Technology.

232. Molenaar, I. (24-11-2011). *It's all about metacognitive activities; Computerized scaffolding of self-regulated learning.* Amsterdam: University of Amsterdam.

231. Brouwer, P. (15-11-2011). *Collaboration in teacher teams.* Utrecht: Utrecht University.

230. Favier, T.T. (31-10-2011). *Geographic information systems in inquiry-based secondary geography education: Theory and practice.* Amsterdam: VU University Amsterdam.

229. Beausaert, A.J. (19-10-2011). *The use of personal developments plans in the workplace. Effects, purposes and supporting conditions.* Maastricht: Maastricht University

228. Kolovou, A. (04-07-2011). *Mathematical problem solving in primary school.* Utrecht: Utrecht University.

227. Schaap, H. (24-06-2011). *Students' personal professional theories in vocational education: Developing a knowledge base.* Utrecht: Utrecht University.

226. Jossberger, H. (24-06-2011). *Towards self-regulated learning in vocational education: Difficulties and opportunities.* Heerlen: Open University of the Netherlands.

225. Dobber, M. (21-06-2011). *Collaboration in groups during teacher education.* Leiden: Leiden University.

224. Van Blankenstein, F.M. (18-05-2011). *Elaboration during problem-based small group discussion: A new approach to study collaborative learning.* Maastricht: Maastricht University.

223. Min-Leliveld, M.J. (18-05-2011). *Supporting medical teachers' learning: Characteristics of effective instructional development.* Leiden: Leiden University.

222. Fastré, G. (11-03-2011). *Improving sustainable assessment skills in vocational education.* Heerlen: Open University of the Netherlands.

221. Slof, B. (28-01-2011). *Representational scripting for carrying out complex learning tasks.* Utrecht: Utrecht University.

220. Bruin-Muurling, G. (21-12-2010). *The development of proficiency in the fraction domain: Affordances and constraints in the curriculum.* Eindhoven: Eindhoven University of Technology.

219. Kostons, D.D.N.M. (05-11-2010). *On the role of self-assessment and task-selection skills in self-regulated learning.* Heerlen: Open University of the Netherlands.

218. Vos, M.A.J. (30-09-2010). *Interaction between teachers and teaching materials: On the implementation of context-based chemistry education.* Eindhoven: Eindhoven University of Technology.

217. Bonestroo, W.J. (24-09-2010). *Planning with graphical overview: Effects of support tools on self-regulated learning.* Enschede: University of Twente.

216. Groenier, M. (10-09-2010). *The decisive moment: Making diagnostic decisions and designing treatments.* Enschede: University of Twente.

215. De Bakker, G.M. (08-09-2010). *Allocated online reciprocal peer support as a candidate for decreasing the tutoring load of teachers.* Eindhoven: Eindhoven University of Technology.

214. Endedijk, M.D. (02-07-2010). *Student teachers' self-regulated learning.* Utrecht: Utrecht University.

213. Kessels, C.C. (30-06-2010). *The influence of induction programs on beginning teachers' well-being and professional development.* Leiden: Leiden University.

212.Duijnhouwer, H. (04-06-2010). *Feedback effects on students' writing motivation, process, and performance.* Utrecht: Utrecht University.

211. Moolenaar, N.M. (01-06-2010). *Ties with potential: Nature, antecedents, and consequences of social networks in school teams.* Amsterdam: University of Amsterdam.

210. Mittendorff, K. M. (12-03-2010). *Career conversations in senior secondary vocational education*. Eindhoven: Eindhoven University of Technology.

209. Platteel, T. (11-02-2010). *Knowledge development of secondary school L1 teachers on concept-context rich education in an action-research setting*. Leiden: Leiden University.

208. Koopman, M. (11-02-2010). *Students' goal orientations, information processing strategies and knowledge development in competence-based pre-vocational secondary education*. Eindhoven: Eindhoven University of Technology.

207. Zitter, I.I. (04-02-2010). *Designing for learning: Studying learning environments in higher professional education from a design perspective*. Utrecht: Utrecht University.