The influence of school size, leadership, evaluation, and time on student outcomes

Four reviews and meta-analyses

Maria Hendriks

THE INFLUENCE OF SCHOOL SIZE, LEADERSHIP, EVALUATION, AND TIME ON STUDENT OUTCOMES

FOUR REVIEWS AND META-ANALYSES

Maria A. Hendriks

Hendriks, M.A.
The influence of school size, leadership, evaluation, and time on student outcomes:
four reviews and meta-analyses

Thesis, University of Twente, 2014

© 2014, Maria Hendriks

ISBN: 978-90-365-3800-8
Doi: 10.3990/1.978-90-365-3800-8

Printed by Ipskamp Drukkers B.V., Enschede, the Netherlands

Painting cover by Coos Vroom and lay out cover by Marc Otte

THE INFLUENCE OF SCHOOL SIZE, LEADERSHIP, EVALUATION, AND TIME ON STUDENT OUTCOMES

FOUR REVIEWS AND META-ANALYSES

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof.dr. H. Brinksma,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op woensdag 3 december 2014 om 14.45 uur

door

Maria Antonia Hendriks
geboren op 11 juli 1959
te Waalwijk

Dit proefschrift is goedgekeurd door de promotor Prof.dr. J. Scheerens.

Samenstelling promotiecommissie

Voorzitter:	Prof.dr.ir. A.J. Mouthaan	Universiteit Twente
Promotor:	Prof.dr. J. Scheerens	Universiteit Twente
Leden:	Prof.dr. B.P.M. Creemers	Rijksuniversiteit Groningen
	Prof.dr. C.A.W. Glas	Universiteit Twente
	Prof.dr. F.J.G. Janssens	Universiteit Twente
	Prof.dr. J.W.M. Kessels	Universiteit Twente
	Prof.dr. P. Van Petegem	Universiteit Antwerpen

Contents

1

Introduction

Context of the Dissertation Study

The orientation of this dissertation originated from the former research program "Effectiveness of School and Training Organizations", of the Department of School Organization and Management (O&M) of the Faculty of Behavioral Sciences of the University of Twente. This program was led by Prof. Dr. Jaap Scheerens and the central research questions were:

- which characteristics of school and training organizations are indicative of high productivity and effectiveness of educational and training provisions?
- which models and theories can explain the operation of these conditions?

Since its start in 1989 one of the main strands of research in this research program was indicated as "foundational", with the aim to establish key concepts and periodically review the existing research evidence. Research reviews and quantitative meta-analyses were the main approaches that were used to accomplish this. Major publications in this area are Scheerens and Bosker (1997), Scheerens, Seidel, Witziers, Hendriks and Doornekamp (2005), Scheerens, Luyten, Steen and Luyten-de Thouars (2007) and Witziers, Bosker and Krüger (2003).

The current dissertation builds on these previous reviews and meta-analyses, focusing on key constructs representing school and instructional factors expected to improve student outcomes. The choice of variables was also determined by funding opportunities. The reviews and meta-analyses on School Size and Learning Time in Schools and Homework were funded by the Netherlands Organisation for Scientific Research, NWO, the study on School Leadership was funded by the Directorate of Knowledge of the Ministry of Education, Culture and Science while the study on Evaluation was funded by the University of Twente. The contents of this dissertation was an important part of three book publications that appeared in 2012, 2013, and 2014 addressing respectively School Leadership Effects, Effectiveness of Time Investments in Education, and School Size Effects Revisited (Scheerens, 2012; Scheerens, 2014a; Luyten, Hendriks & Scheerens, 2014).

School Effectiveness Research

School effectiveness research addresses the question why and how some schools are more effective than others when the differences in achievement cannot be attributed to student intake and educational background characteristics. A main aim is to identify and investigate those malleable conditions at different levels –classroom, school and above school– that can directly or indirectly explain the differences in the learning outcomes of students (Creemers & Kyriakides, 2008; Reynolds, Sammons, De Fraine, Van Damme, Townsend, Teddlie & Stringfield, 2014; Scheerens, 2013).

School effectiveness research emerged in the 1970s as a response to the work of Coleman et al. (1966) and Jencks et al. (1972) who stated that 'schools and schooling did not make a difference'. After a first phase in which school effectiveness research focussed on showing that 'school matters', effectiveness studies tried to open the 'black box' of schooling in order to explore the reasons why schools had their different effects. In this phase researchers were concerned with identifying characteristics of schools and teachers that might explain the differences in educational outcomes. The studies resulted in consistent and partly overlapping lists of effectiveness enhancing factors (Reynolds et al., 2014). The first and well-known list is the five factor model (Edmonds, 1979) in which effective schools were characterized by strong educational leadership, high expectations of student achievement, an emphasis on basic skills, a safe and orderly climate and frequent evaluation of student progress. These factors appear to be still valid today as would appear from recent narrative reviews and meta-analyses (see e.g. Kyriakides, Creemers, Antoniou & Demetriou, 2010; Scheerens, 2013, 2014b).

School effectiveness research stems from different research traditions and disciplinary perspectives, including (in)equality of education (sociological perspective), educational production functions (economical perspective), evaluation compensatory programs, effective schools and teacher and instruction effectiveness (psychological perspective). A sixth research orientation, system level effectiveness, is emerging. The various traditions each concentrated on different types of conditions that were assumed to be associated with positive educational outcomes and different organizational levels (school, classroom and above school level) (Creemers & Kyriakides, 2008; Scheerens, 2013). During the past two decades researchers have taken a more comprehensive view on school effectiveness. Integrated multilevel models of school effectiveness were introduced in which the key effectiveness enhancing conditions from each research tradition were included, each on the appropriate level of functioning. Examples of these comprehensive models are those by Scheerens (1992), Stringfield and Slavin (1992), Creemers (1994) and more recently the dynamic model of educational effectiveness by Creemers and Kyriakides (2008).

Common characteristics of these models are that they take into account multiple factors of effectiveness that operate at different levels. Effectiveness enhancing conditions at the classroom or teaching and learning level are the core of the comprehensive models, with the conditions at classroom level usually organized according to the Carroll model of schooling (Carroll, 1963). Important variables in the Carroll model are time for learning, opportunity to learn and classroom instruction. School level conditions are seen as facilitating conditions of effective classroom conditions, but multilevel modelling also shows that school and classroom factors can also influence each other reciprocally (see e.g. Bosker & Scheerens, 1994). What is more, some variables (e.g. monitoring pupil’s progress or time for learning) are meaningful both at class- and school level.

In the dynamic model of educational effectiveness the functioning of each factor is seen from a dynamic and an instrumental perspective (Creemers, Kyriakides & Sammons, 2010). The dynamic approach to educational effectiveness research adds to the comprehensive model a need for longitudinal research in studying development over time, both with regard to the outcomes as also the effectiveness enhancing conditions at student, class, school, and context level. Further characteristics of the dynamic model concern:

- the assumption that the relationships between some effectiveness enhancing conditions and outcomes might be non-linear;
- the need to carefully examine the interrelations between factors operating at the same level;
- the use of different dimensions to define the effectiveness enhancing factors, and;
- the adoption of further outcomes of learning than basic skills in language and math, including affective and psycho-motoric outcomes as well as achievement outcomes that derive from new ways of learning aimed at self-regulated learning and lifelong learning (Creemers & Kyriakides, 2008; Muijs, Kyriakides, Van der Werf, Creemers, Timperley & Earl, 2014; Scheerens, 2013).

The Robustness of the Knowledge Base

From the beginning of school effectiveness research researchers conducted narrative reviews to compile the state of the art knowledge and to identify the factors that matter most (see e.g. Cotton, 1995; Levine & Lezotte, 1990; Sammons, Hillman & Mortimore, 1995; Scheerens, 1992). Recently in two 'state-of-the-art' review studies Reynolds et al. (2014) and Muijs et al. (2014) synthesized the evidence of the research on school effectiveness and teacher effectiveness respectively. Results from these recent reviews show that there is still considerable consensus with regard to the main effectiveness enhancing conditions that also appeared in the earlier reviews, i.e. achievement orientation, time for learning, opportunity to learn, classroom management, structuring and scaffolding of instruction, feedback, effective leadership and monitoring progress.

Later research has added important specification and further differentiation of school level variables, as well as more emphasis on classroom level instructional variables, with recently also interest into contextual influences such as the role of local authorities and school districts at above school level and the influence of policies and institutional arrangements at system level (Sammons, 2012; Scheerens, 2013).

Although there thus seems to be considerable consensus regarding the general factors 'that work', the actual operationalization of each of the effectiveness enhancing conditions lacks agreement. The variety of operational definitions used in the primary studies and the tendency to constantly re-invent the wheel in defining key-variables and measurement instruments impedes the development of a robust knowledge base (Muijs, 2012; Scheerens, 2014b).

Moreover, results from meta-analyses (see e.g. Creemers & Kyriakides, 2008; Hattie, 2009; Kyriakides, Christoforiu & Charalambous, 2013; Scheerens & Bosker, 1997, Scheerens et al., 2007; Seidel & Shavelson, 2007) show less consensus as well, as far as the magnitude of the average effect size of the relationship with an effectiveness enhancing factor and student outcomes is concerned. While some meta-analyses (i.e. Hattie, 2009; Kyriakides et al., 2013) report average effect sizes that are medium according to established scientific standards, the average effects for the same effectiveness enhancing factor reported in other meta-analyses are relatively small. These differences might be due to methods employed in the meta-analyses as well to methodological flaws in both the original studies as well as the meta-analyses (see e.g. Kohn, 2006; Scheerens, 2013, 2014b). The small effects reported by Seidel and Shavelson (2007) e.g. might be explained by the fact that these authors applied more strict inclusion criteria than others as they only included studies that had controlled for student prerequisites. Next, the meta-analyses differed considerably in the amount of studies included and the countries in which the studies were employed. Reported average effects sizes might be higher if the studies included are mainly conducted in the USA, Great Britain and Australia as in these countries the variance might be larger in both effectiveness enhancing variables and outcomes.

The considerable variability in effect sizes, however, gives reason to be cautious in interpreting the strength of the educational effectiveness knowledge base.

Meta-Analysis

Meta-analysis summarizes statistical results from a range of independent studies that address a related research question. Meta-analysis is sometimes used as a synonym for systematic review. However, the term systematic review is usually used for the systematic search, retrieval, and assessment of research studies, while the term meta-analysis is used to describe the quantitative procedures to statistically combine the results of studies (Cooper, Hedges & Valentine, 2009).

Before meta-analysis became more common in the 1980s studies were summarized in a narrative review or combined in the so-called vote counting technique. Vote counting basically consists of the counting the number of positive and negative significant and non-significant associations. Vote counting, however, does not take into account the strength of the relationship (i.e. how large the effect size is), neither does it incorporate the sample size into the vote. Therefore vote counting is seen as a “next best” solution to meta-analysis. In the reviews and meta-analyses included in this dissertation study the main reason to use the vote count method was that a sizeable number of studies did not provide sufficient information to permit calculation of an effect size. In order to not throw away the information from these studies the less demanding vote count procedure was applied as well.

Compared to traditional review procedures one of the most distinguished features of meta-analysis is the conversion of individual study results in a common metric, an effect size statistic. By standardizing effect sizes of individual studies it is possible to compare across different studies as well as to integrate results. The first stage in a meta-analysis is usually to establish an average effect size and an estimate of the statistical significance of the relationship (a confidence interval). Often, meta-analysts are even more interested in determining how the primary studies differ from each other. A homogeneity test of effect sizes is then applied to show whether there are systematic differences between studies. And, if there appears to be variability, and in most cases there is, it is needed to run moderator analysis that can help to determine the features of the studies that may explain these differences. Various models (fixed effects, random effects and multilevel models) have been developed to examine the degree to which the variability in effect sizes could be attributed to specific study characteristics.

Early meta-analyses were based on a fixed-effects model which assumes that all studies in the analysis estimate the same underlying true effect size and the variability between effect size estimates is due to sampling error alone (Borenstein, Hedges, Higgins & Rothstein, 2010). In reality this is rarely the case (Field & Gillet, 2010). More recently therefore, researchers have argued for a random effects model. The random effects model allows that there may be a distribution of true effect sizes. In the random effects model the amount of variance is assumed to reflect both sampling error plus variability assumed to be randomly distributed in the population of effects.

An important assumption of both the fixed effects and random effects model is the assumption of statistical independence (Cooper et al., 2009; Lipsey & Wilson, 2001). This implies that if a study reports multiple effect sizes, only one effect size per study could be considered. Also, meta-analysis can violate the assumption of independence when more than one treatment group or sample is included in the same study. Multilevel meta-analysis techniques can be applied to account for such dependencies, or correlations, within the studies (Hox, 2002). A further major advantage of the multilevel approach compared to the fixed effects and random effects models is its flexibility in modelling the data, e.g. when one has multiple moderator variables or when one wants to accommodate for multiple outcome measures (Hox, 2002; Raudenbush & Bryk, 2002).

Overview of the Contents

As indicated in the above the dissertation reports on four reviews and meta-analyses focused at the effects of School Size, School Leadership, Evaluation and Learning Time on student outcomes. The four reviews and meta-analyses explore factors at different levels of the conceptual school effectiveness models. While factors as School Size and School Leadership usually have meaning at school level, Evaluation and Learning Time can be conceptualized at school and classroom level. What is more, depending on the available data, different methods for review and meta-analysis were applied to integrate the findings of individual studies and to draw conclusions about the impact of the four school and classroom factors concerned.

School Size

In the research on school size effects two main perspectives can be distinguished: on the one hand the effectiveness perspective, in which research is focused on the impact of school size on educational outcomes, and on the other hand, the efficiency perspective in which research is focused on the cost effectiveness of school size. A third perspective is the embedding of school size in multilevel school effectiveness models. In conceptual multilevel school effectiveness models school size usually is included as context variable at school level and not immediately seen as one of the malleable variables that might have a positive impact on achievement. Gaining a better insight into the other preconditions and intermediate school and instruction characteristics that facilitate or impede the effects of school size on educational outcomes is a third perspective in the study in Chapter 2. The main research questions addressed in Chapter 2 are:

1. What is the impact of school size on various cognitive and non-cognitive outcomes and school organizational outcome variables?
2. What is the "state of the art" of the empirical research on economies of size?
3. What is the direct and indirect impact of school size, conditioned by other school context variables on student performance (where indirect effects are perceived as influencing through intermediate school and instruction characteristics)?

To answer the first and third question the impact of school size on a variety on student, teacher, parents' and school organizational outcome variables was investigated. In the study school organization variables are considered as a desirable end in itself, but also as intermediate variables conducive to high academic performance and positive student and teacher attitudes. To answer the second question, costs was included as a dependent variable.

The study summarizes the results of 84 empirical studies on the impact of school size on various student, teacher and school organizational outcomes. A vote count procedure was applied as well as a narrative review, providing more in-depth information on some of the studies.

School Leadership

Earlier reviews and meta-analyses of leadership effects were based on 'direct' effect models of leadership on student performance outcomes. Basically, simple correlations between leadership characteristics and student achievement, sometimes adjusted for student background characteristics, were at the focus of these reviews.

Chapter 3 focuses on leadership effect studies that employed an indirect effect model. These mediated or indirect effect models hypothesize school leaders to achieve their effect on school performance not only through a direct effect from school leadership to student achievement, but also through intermediate variables such as school organization and school culture.

The main research questions addressed in Chapter 3 are:

1. What is the total (direct and indirect) effect of school leadership on student achievement?
2. What are the most promising paths and intermediate variables in indirect effect models that study the impact of school leadership on student achievement?

The study summarizes the results of 15 leadership effect studies that used indirect-effect models. A quantitative meta-analysis was applied as well as a narrative review, proving information on the intermediary variables that could play a role in explaining indirect school leadership effects.

Evaluation

One of the five factors Edmonds (1979) drew forward on the basis of school effectiveness research was frequent monitoring of student performance. So from the early days of effective schools' research onwards evaluation and assessment has been mentioned as part of a limited set of effectiveness enhancing conditions and this has not changed. Evaluation and assessment remain prominently present in recent reviews of the literature.

The main research question of the study presented in Chapter 4 was: "What is the impact of evaluation and assessment on student achievement at both school and classroom level?

The meta-analysis included 7 studies on evaluation at school level, 14 studies on evaluation at class level and 6 studies examining the impact of assessment. A random effects model was applied to calculate the weighted mean effect sizes. A vote count procedure was applied as well to permit the inclusion of studies that did not provide sufficient information to calculate an effect size.

Learning Time in Schools and Homework

Time for schooling and teaching is considered one of the key variables to improve educational outcomes and the quality of schooling. The underlying notion, namely that good schooling and teaching depends on the "exposure" of students is clear and plausible.

In earlier meta-analyses on the effect of learning time in school and homework on student achievement, a broad range of different operational definitions of time was used in the primary studies. As the effects of this mixture of different specifications were thrown together in the meta-analyses, the findings could only be interpreted as a general overall effect of time. In addition to the general effect of time the meta-analysis presented in Chapter 5 also addresses the differential effects of facets of learning time and homework. The second aim of the meta-analysis was to address potential moderators of the effects of time for learning and homework.

The meta-analysis included 12 studies on learning time in schools, and 23 studies for homework. A multilevel meta-analysis was conducted based on the approach outlined by Hox (2002). A random effects model was fitted. Moderator analyses were conducted to examine the degree to which the relationship between learning time or homework on the one hand and student achievement on the other could be attributed to specific sample or study characteristics.

In the final chapter the main results of each chapter are reviewed, and general issues resulting from all four chapters are discussed.

References

Borenstein, B., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*, 97-111. doi:10.1002/jrsm.12

Bosker, R. J., & Scheerens, J. (1994). Alternative models of school effectiveness put to the test. In R. J. Bosker, B. P.M . Creemers & J. Scheerens (Eds.), *Conceptual and methodological advances in educational effectiveness research* (pp. 159-180). Special issue of the International *Journal of Educational Research, 21*(2).

Carroll, J. B. (1963). A model of school learning. *Teachers College Record*, *64*(8), 722-733.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F.,& York, R. (1966). *Equality of educational opportunity.* Washington D.C.: U.S. Government Printing Office.

Cooper, L., Hedges, L. V., &Valentine, J.C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage.

Cotton, K. (1995). *Effective schooling practices: A research synthesis. 1995 Update*. School Improvement Research Series. Northwest Regional Educational Laboratory.

Creemers, B. P. M. (1994). *The effective classroom.* London: Cassell.

Creemers, B. P. M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness.* London and New York: Routledge.

Creemers, B. P. M., Kyriakides, L., & Sammons, P. (2010). *Methodological advances in school effectiveness research.* London: Routledge.

Edmonds, R. (1979). Effective schools for the urban poor. *Educational Leadership, 37*, 15-27.

Field, A. P., & Gillett, R. (2010). Expert tutorial. How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *63*, 665-694. doi:10.1348/000711010X502733

Hox, J. (2002). *Multilevel analysis techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Jencks, C. S., Smith, M., Ackland, H., Bane, M. J., Cohen, D., Gintis, H., Heyns, B., & Michelson, S. (1972). *Inequality: A reassessment of the effect of the family and schooling in America.* New York, NY: Basic Books.

Kohn, A. (2006). Abusing research: The study of homework and other examples. *Phi Delta Kappan, 88*(1), 8-22.

Kyriakides, L., Creemers, B., Antoniou, P., & Demetriou, D. (2010). A synthesis of studies searching for school factors: implications for theory and research. *British Educational Research Journal*, *36*, 807-830. doi:10.1080/01411920903165603

Kyriakides, L., Christoforou, C., & Charalambous, C. L. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education*, *36*, 143-152. doi:10.1016/j.tate.2013.07.010

Levine, D. U., & Lezotte, L. W. (1990). *Unusually effective schools: A review and analysis of research and practice.* Madison, WI: National Center for Effective Schools Research and Development.

Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Luyten, H., Hendriks, M. A., & Scheerens, J. (Eds.) (2014). *School size effects revisited* (SpringerBriefs in Education). Cham: Springer.

Muijs, D. (2012). Methodological change in educational effectiveness research. In C. Chapman, P. Armstrong, A. Harris, D. Muijs, D. Reynolds & P. Sammons (Eds.), *School effectiveness and improvement research, policy and practice* (pp. 58-66). London and New York: Routledge.

Muijs, D., Kyriakides, L., Van der Werf, G., Creemers, B. P. M, Timperley, H., & Earl, L. (2014). State of the art – teacher effectiveness and professional learning. *School Effectiveness and School Improvement, 25*, 231-256. doi:10.1080/09243453.2014.885451

Raudenbush, S. W., & Bryk, A.S. (2002). *Hierarchical linear modelling* (2nd ed.). Thousand Oaks, CA: Sage.

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J. Townsend, T., Teddlie, Ch., & Stringfield, S. (2014). Educational effectiveness research (EER): a state-of-the-art review. *School Effectiveness and School Improvement, 25*, 197-230. doi:10.1080/09243453.2014.885450

Sammons, P. (2012). Methodological issues and new trends in educational effectiveness research. In C. Chapman, P. Armstrong, A. Harris, D. Muijs, D. Reynolds & P. Sammons (Eds.), *School effectiveness and improvement research, policy and practice* (pp. 9-26). London and New York: Routledge.

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

Scheerens, J. (1992). *Effective schooling, research, theory and practice*. London: Cassell.

Scheerens, J. (Ed.) (2012). *School leadership effects revisited. Review and meta-analysis of empirical studies* (Springer Briefs in Education). Dordrecht: Springer.

Scheerens, J. (2013). What is effective schooling? A review of current though and practice. Retrieved from www.ibo.org/research/resources

Scheerens, J. (Ed.) (2014a). *Effectiveness of time investments in education* (SpringerBriefs in Education). Cham: Springer.

Scheerens, J. (2014b). School, teaching, and system effectiveness: some comments on three state-of-the-art reviews. *School Effectiveness and School Improvement, 25*, 282-290. doi:10.1080/09243453.2014.885453

Scheerens, J., & Bosker, R. (1997). The foundations of educational effectiveness. Oxford: Pergamon.

Scheerens, J., Luyten, H., Steen, R., & Luyten-de Thouars, Y. (2007). *Review and meta-analyses of school and teaching effectiveness.* Enschede: University of Twente, Department of Educational Organisation and Management.

Scheerens, J., Seidel, T., Witziers, B., Hendriks, M., & Doornekamp, G. (2005). *Positioning and validating the supervision framework.* Enschede: University of Twente, Department of Educational Organisation and Management.

Seidel, T., & Shavelson, R.J. (2007). Teaching effectiveness research in the past decade: the role of theory and research design in disentangling meta-analysis results. *Review of Educational Research,* 77, 454-499. doi:10.3102/0034654307310317

Stringfield, S. C., & Slavin, R. E. (1992). A hierarchical longitudinal model for elementary school effects. In B. P. M. Creemers & G. J. Reezigt (Eds.), *Evaluation of educational effectiveness* (pp. 35-68). Groningen: ICO.

Witziers, B., Bosker, R. J., & Krüger, M. L. (2003). Educational leadership and student achievement: the elusive search for an association. *Educational Administration Quarterly, 39*, 398-425. doi:10.1177/0013161X03253411

2

School size effects; A synthesis of studies published between 1990 and 2012[1]

[1] This chapter is based on Hendriks, M. A. (2014). Research synthesis of studies published between 1990 and 2012. In H. Luyten, M. A. Hendriks & J. Scheerens (Eds.), *School size effects revisited* (SpringerBriefs in Education) (pp. 41-175). Cham: Springer.

Abstract

Size of school organizations has received considerable attention in education policy and scale is expected to have an impact on the social and affective dimensions of schooling. This review synthesis summarizes the results of 84 empirical studies on the impact of school size on various student, teacher and school organizational outcomes. A vote count procedure was applied as well as a narrative review, providing more in-depth information on some of the studies. The results of the review challenge some of the beliefs about small school size, but are in line with those from earlier reviews. With regard to academic achievement no clear results are found as the majority of reported school size effects failed to reach statistical significance. For non-cognitive outcomes like safety and school attendance the review revealed mixed results. When social cohesion and student, teacher or parent participation were the outcome measures the findings were in the expected direction and clearly showed a positive impact of smaller schools. Just a few studies addressed the indirect effects of school size. Future research therefore should not only aim at the outcomes of school size but try to clarify the preconditions and intermediating school and instructional effects of school size as well and so try to open the black box of positive, negative, curvilinear and non-significant school size effects found in this review study.

Introduction

Size of school organizations is a recurrent theme in educational policy. For a long period of time education policy in countries like the United States and the Netherlands has been focused on stimulating scaling-up. The expectation was that larger schools would be cost-effective and beneficial to the quality of education and the education career opportunities for pupils. Within larger institutions it was assumed that pupils do have wider curricular and extracurricular choice and better transfer opportunities to other programs. Moreover, larger schools provide more opportunities for professionalization and specialization of staff and have lower per-pupil costs. On the other hand, during the past years, interest in side effects and potential risks of scaling-up has simultaneously increased. The undesirable effects are related to limitations in the freedom of choice of students and parents, to increased managerial overhead and to diminishing social cohesion within the institutions (Onderwijsraad, 2005). In smaller educational institutions it might be easier to create a more personalized learning environment, and there are better chances of higher commitment, interaction and participation by students, parents and teachers (see e.g. Cotton, 2001; Newman et al., 2006). In the United States these claims led to many reforms, where traditional large high schools were converted into smaller more personal schools, mainly supported by institutions such as the Bill and Melinda Gates Foundation (Kahne, Sporte, De La Torre & Easton, 2008; NWO, 2011). In other countries the same debates with regard to scale are visible (NWO, 2011). At the same time the research literature has not yet produced consistent empirical evidence about the impact of school size on educational outcomes (see e.g. prior reviews by Andrews, Duncombe & Yinger, 2002; Leithwood & Jantzi, 2009; Newman et al., 2006) although the evidence seems to be somewhat stronger for non-cognitive than for cognitive outcomes. Perceptions on school climate and social cohesion are generally found more positive in smaller schools. Also different optimum school sizes are found depending on the country in which the study was conducted, the level of schooling the study focused on (e.g. primary or secondary education) and the socio economic background of the student population. Less is known about the indirect effects of school size, i.e. the intermediate school organization and teaching and learning variables such as a more personalized climate or a more focused curriculum, which are directly affected by changes in school size and which in their term may affect educational outcomes (NWO, 2011).

In the research on school size effects two main perspectives can be distinguished. On the one hand there is the basic question of the impact of school size on educational outcomes, which we consider as the effectiveness perspective. On the other hand, research is focused on the cost effectiveness of school size, which is considered the efficiency perspective. A third perspective, which can be seen as a further elaboration of the effectiveness perspective, is the embedding of school size in multilevel school effectiveness models. In conceptual multilevel school effectiveness models (see e.g. Scheerens, 1992; Scheerens & Bosker, 1997) school size usually is included as a context variable at school level and not immediately seen as one of the malleable variables that might have a positive impact on achievement. Gaining a better insight into the other preconditions and intermediate school and instruction characteristics that facilitate or impede the effects of school size on educational outcomes is a third perspective (Scheerens, Hendriks & Luyten, 2014a).

In this chapter the results of a research synthesis of the effects on school size on various outcome variables are presented. The present review builds on an earlier “quick scan” on the impact of secondary school size on achievement, social cohesion, school safety and involvement conducted for the Dutch Ministry of Education and Sciences in 2008 (Hendriks, Scheerens & Steen, 2008). The research synthesis seeks to answer the following questions:

- What is the impact of school size on various cognitive and non-cognitive outcomes?
- What is the “state of the art” of the empirical research on economies of size?
- What is the direct and indirect impact of school size, conditioned by other school context variables on student performance (where indirect effects are perceived as influencing through intermediate school and instruction characteristics)?

To answer the first and third question the impact of school size of variety of student, teacher, parents’ and school organizational outcome variables was investigated. A distinction is made between different outcome variables, i.e. cognitive and non-cognitive outcome variables, and school organization variables. Cognitive outcomes refer to student achievement. The non-cognitive outcome variables included in the review relate both to students (attitudes towards school and learning, participation, safety, engagement, absence and drop-out), to parents (participation) and teachers (satisfaction, commitment and efficacy).

School organization variables relate to safety, to involvement of students, teachers and parents, as well as to other aspects of the internal organization of the school, including classroom practices (i.e. aspects of teaching and learning). In the review school organization variables are considered as a desirable end in itself, but also as intermediate variables conducive to high academic performance and positive student and teacher attitudes. To answer the second question, costs was included as a dependent variable.

In the current review it was not possible to apply a quantitative meta-analysis in which effect sizes are combined statistically. One reason was that many empirical studies did not provide sufficient information to permit the calculation of an effect size estimate. What is more, in many cases the relationship of school size and a dependent variable is not modeled as a linear relationship. Instead a log-linear or quadratic relationship is examined or different categories of school size are compared, of which the number and distribution of sizes over categories varied between studies.

Therefore we used the so-called vote count technique, which basically consists of counting the number of positive and negative statistically significant and non-significant associations. This technique has limitations, as will be documented in more detail when presenting the analyses. In this chapter the results of the vote counts as well as a narrative review, providing more in-depth information of a number of the studies, are presented.

Method

Search Strategy and Selection Criteria

A computer assisted literature search procedure was conducted to find empirical studies that investigated the impact of school size on a wide array of student outcomes (such as achievement, cohesion, safety, involvement, participation, attendance, drop-out and costs). Literature searches of the electronic databases Web of science (www.isiknowledge.com), Scopus (www.scopus.com), ERIC, Psycinfo (provided through Ebscohost) and Picarta were conducted to identify eligible studies. Search terms included key terms used in the meta-analysis by Hendriks, Scheerens and Steen (2008), i.e. (a) “school size”, “small* schools”, “larg* schools”, (b) effectiveness, achievement, (c) cohesion, peer*, climate, communit*, “peer relationship”, “student teacher relationship”, (d) safe*, violence, security, (e) influenc*, involvement, participation, (f) truancy, “drop out”, attendance and (g) costs. In the search the key terms of the first group were combined with the key terms of each other group separately. We used the limiters publication date January 1990 - October 2012 and peer reviewed (ERIC only) to restrict our search.

The initial search in the databases yielded 1984 references and resulted in 875 unique studies after removing duplicate publications. The titles and abstracts of these publications were screened to determine whether the study met the following criteria:

- The study had to include a variable measuring individual school size. Studies investigating schools-within-schools or studies examining size at the school district level were not included in the review. Studies were also excluded if school size was measured as grade or cohort enrolment or the number of teachers in the school.

- The dependent variable of the study had to be one or more of: student attainment and progress, student behavior and attitudes, teacher behavior and attitudes, school organizational practices and teaching and learning, and; economic costs.
- The study had to focus on primary or secondary education (for students aged 6-18). Studies that focused on preschool, kindergarten or on postsecondary education were excluded.
- The study had to be conducted in mainstream education. Studies containing specific samples of students in regular schools (such as students with learning, physical, emotional, or behavioral disabilities) or studies conducted in schools for special education were excluded from the review.
- The study had to be published or reported no earlier than January 1990 and before December 2012.
- The study had to be written in English, German or Dutch.
- The study had to have estimated in some way the relationship between school size and one or more of the outcome variables. Studies had to report original data and outcomes. Existing reviews of the literature were excluded from the review.
- When cognitive achievement was the outcome variable, studies had to control for a measure of students' background, such as prior cognitive achievement and/or socio-economic status (SES).

After this first selection, 314 studies left for the full text review phase. In addition recent reviews on school size (i.e. Andrews et al., 2002; Hendriks et al., 2008; Leithwood & Jantzi, 2009; Newman et al., 2006) as well as references from the literature review sections from the obtained publication were examined to find additional publications. A cut-off date for obtaining publications was set at 31 December 2012.

The full text review phase resulted in 84 publications covering the period 1990-2012 admitted to the review and fully coded in the coding phase. Because our review is more recent we were able to provide a more up-to-date overview of the empirical evidence on school size. In this review we included 73 studies not covered in the review by Newman et al. (2006) and 60 studies not incorporated in the review by Leithwood & Jantzi (2009).

The data were extracted by one of two reviewers and confirmatory data extraction was carried out by a second reviewer.

Coding Procedure

Lipsey and Wilson (2001) define two levels at which the data of the study should be coded: the study level and the level of an effect size estimate. The authors define a study as "a set of data collected under a single research plan from a designated sample of respondents" (Lipsey & Wilson, p. 76). A study may contain different samples, when the same research is conducted on different samples of participants (e.g. when students are sampled in different grades, cohorts of students or students in different stages of schooling -primary or secondary-) or when students are sampled in different countries. An estimate is an effect size, calculated for a quantitative relationship between an independent and dependent variable. As a study may include different measurements of the *independent* variable (school size), as well as different measures of the *dependent* variable (such as e.g. different outcome measures (achievement, engagement, drop-out), different achievement tests covering different domains of subject matter(e.g. language or math), measurement as different point in time (e.g. learning gain after two and four years), a study may yield many effect sizes, each estimate different from the others with regard to some of its details.

The studies selected between 1990 and 2012 were coded by the researchers applying the same coding procedure as used by Scheerens, Luyten, Steen and Luyten-de Thouars (2007). The coding form included five different sections: report and study identification, characteristics of the independent (school size) variable(s) measured, sample characteristics, study characteristics and school size effects (effect sizes).

The report and study identification section recorded the author(s), the title and the year of the publication.

The section with characteristics of the explanatory variable(s) measured coded the operational definition of the school size variable(s) used in the study (in all studies referring to a measure of total number of students attending a school) as well as the way in which the relationship between size and outcomes was modeled in the study: either linear or transformed to its logarithm (size measured as a continuous variable), quadratic (estimating both linear and quadratic coefficients) or comparing different size categories.

The sample characteristics section recorded the study setting and participants. For study setting the country or countries in which the study was conducted were coded. With regard to participants, the stage of schooling (primary or secondary level) the sample referred to was coded as well as the grade or age level(s) of the students the sample focused on. The number of schools, classes and students included in the sample were recorded as well.

The study characteristics section coded the research design chosen, the statistical techniques conducted and the model specification. For the type of research design we coded whether the study applied a quasi-experimental or experimental research design and whether or not a correlational survey design was used. The studies were further categorized according to the statistical techniques conducted to investigate the association between school size and achievement. The following main categories were employed: analysis of variance, Pearson correlation analysis, (logistic) regression analysis, path analysis/LISREL/ SEM, multi-level analysis as well as specific methods for economic analyses such as two stage least-square regression. We also coded whether the study accounted for covariates at the student level, i.e. if the study controlled for prior achievement, ability and/or student social background.

Finally, the school size effects section recorded the effects sizes, either taken directly from the selected publications or calculated. The effect sizes were coded as reflecting the types of outcome variables distinguished in the review (i.e. achievement, students' and teachers' attitudes to school, students', teachers' and parents' participation, safety, attendance, absenteeism, truancy and drop out, school organization and teaching and learning, and costs). With regard to achievement, four groups of academic subjects were distinguished in the coding: language, mathematics, science and other subjects.

"Vote Counting" Procedure

Vote counting comes down to counting the number of positive significant, negative significant and non-significant associations between an independent variable and a specific dependent variable of interest from a given set of studies at a specified significance level, in this case school size and different outcome measures (Bushman & Wang, 2009). We used a significance level of α=.05. When multiple effect size estimates were reported in a study, each effect was individually included in the vote counts.

The vote counting procedure has been criticized on several grounds (Borenstein, Hedges, Higgins & Rothstein, 2009; Bushman, 1994; Bushman & Wang, 2009; Scheerens, Seidel, Witziers, Hendriks & Doornekamp, 2005). It does not incorporate sample size into the vote. As sample sizes increase, the probability of obtaining statistically significant results increases. Next, the vote counting procedure does not allow the researcher to determine which treatment is the best in an absolute sense as it does not provide an effect size estimate. Finally, when multiple effects are reported in a study, such a study has a larger influence on the results of the vote count procedure than a study where only one effect is reported. Therefore vote counting is seen as a "next best" solution, which we choose to apply given the limitations of the set of basic studies, explained in the introduction.

Vote counting procedures were applied for each of the (groups of) dependent variables: student achievement, students' and teachers' attitudes to school, students', teachers' and parents' participation, safety, attendance, absenteeism, truancy and drop out, school organization and teaching and learning, and costs.

Table 2.1 gives an overview of the studies, samples and estimates included in the vote counting procedures for each type of outcome variables (i.e. achievement, students' and teachers' attitudes to school, students', teachers' and parents' participation, safety, attendance, absenteeism, truancy and drop out, school organization and teaching and learning, and costs) as well as in total.

Table 2.1
Number of studies, samples and estimates included in the vote-counting procedure for each (group of) dependent variable(s) and in total

	Studies	Samples	Number of significant or non-significant effects
Achievement	46	64	126
Students' and teachers' attitudes to school	14	15	24
Participation	10	10	13
Safety	24	25	54
Attendance, absenteeism and truancy	12	19	23
Drop-out	4	5	5
Other student outcomes	5	6	9
School organization and teaching and learning	4	4	18
Costs	5	5	5
Total	84	107	277

Analysis of Study and Sample Characteristics

So-called "moderator variables" were taken into account to examine the degree to which the relationship between school size on the one hand and an outcome variable on the other would appear to be attributable to specific sample or study characteristics. In the case of vote counts this comes down to providing more specific cross-breaks for the sub-categories of the study characteristics seen as moderators. Due to the low number of samples included in the review for most of the outcome variables (see Table 2.3), analysis of such study and sample characteristics was only applied for those studies and samples that included student achievement or safety as the outcome variable, and in which the relationship between size and outcomes was modeled as a linear or log-linear function. The following types of study and sample characteristics were used in our analyses: sample characteristics as geographical region, the level of schooling (primary, secondary schools), and study characteristics that refer to methodological and statistical aspects, e.g. study design, model specification, whether or not covariates at the student level (SES, cognitive aptitude, prior achievement) or school level (school level SES, urbanicity) are taken into account and whether or not multilevel analysis was employed.

A total of 84 studies and 107 samples were included in the review. Almost three quarter of the studies (i.e. 58 studies) originate from the United States. Seven studies were conducted in the Netherlands, four in the United Kingdom, three in Israel, two in Canada, two in Sweden and one in each of Australia, Hong Kong, Ireland, Italy and Taiwan.

Eighteen studies examined effects of school size in primary education contexts, 53 studies in secondary schools and six studies collected data in primary and secondary schools separately. In three studies a combined sample of primary and secondary schools was used.

Results

Results of studies on school size effects are presented for various outcome variables: academic achievement, social cohesion, participation and commitment of students and teachers, student absence and dropout and other outcome variables. School size effects were also studied with school organizational characteristics and costs as the dependent variable.

Academic Achievement

Evidence about the relationship between school size and academic achievement was derived from 46 studies and 64 samples (yielding in total 126 effect estimates). Twenty studies (22 samples) provided evidence about the relationship between school size and achievement in primary education. Evidence about the effects of school size in secondary education was available from 29 studies (39 samples). In five studies the data were obtained from samples that included students from both levels of schooling. The vast majority of studies (and samples) were conducted in the United States. The other studies originate from Canada (1 sample), Hong Kong (1 sample), the Netherlands (2 samples) and Sweden (2 samples).

More detailed information about the characteristics of the samples and studies that examined the impact of size on student achievement can be found in Table A1.

Table 2.2 shows the results of the total number of negative, non-significant, curvilinear and positive effects found for the associations between school size and cognitive achievement.

Table 2.2

Results of vote counts examining the number of negative, non-significant, curvilinear and positive effects of school size on achievement

			Direction of effect			
	Studies	Samples	-	ns	∩	+
School size measured as a continuous variable	31	46	20	62	0	8
School size squared measured	4	8	0	0	8	0
School size measured as discrete variable (categories)	15	18	3	16	6	3
Total	46	64	23	78	14	11

-	= negatively related with school size
ns	= no significant relation with school size
∩	= optimal school size found
+	= positively related with school size

In this table evidence is presented for all studies in total as well as separately for the three different ways in which school size was measured in the studies: 1) school size measured as a continuous variable usually operationalized as the total number of students attending a school or different sites of a school at a given date, suggesting a linear relationship, 2) school size measured as a quadratic function, seeking evidence for a curvilinear relationship and, 3) school size measured through comparison of different categories. In these latter studies, the evidence reported could show either a linear or curvilinear relationship, on the impact of size categories.

The results of the vote counting show that of 126 effects sizes in total, more than half of the associations (78 effects, 62%) between school size and achievement appeared to be non-significant, 23 estimates (18%) showed negative effects and 11 estimates (9%) positive effects.

School Size Measured as a Continuous Variable

When school size was measured as a continuous variable, in 11 of the 46 samples (20 effects, 22%) a negative relationship between school size and achievement was reported while in 8 samples (8 effect sizes, 9%) it was found that achievement was higher for larger schools (see Table 2.2).

In 15 samples the effects of school size were examined for more than one achievement measure (e.g. in different domains (language or math), or at different points in time). For 14 of these samples the effects found were all in the same direction, thus, either non-significant, positive or negative. The only study that reported mixed results was the study by Fowler & Walberg (1991). In this study five of the achievement measures appeared to be negatively associated with school size; the other eight effects were non-significant.

Besides Fowler & Walberg's study eight other studies (samples) also found negative associations between school size and achievement. In seven of these studies the (weak) negative effects found referred to evidence derived from studies (samples) conducted in primary education (Archibald, 2006; Caldas, 1993; Deller & Rudnicki, 1993; Driscoll, Halcoussis & Svorny, 2003; Heck, 1993; Moe, 2009; Stiefel, Schwartz & Ellen, 2006), while only study conducted in secondary education (Lee & Smith, 1995) reported a negative effect.

On the other hand four of the five studies that found a positive relationship between size and achievement (i.e. achievement went up as school size increased) were conducted in secondary education (Bradley & Taylor, 1998; Foreman-Peck & Foreman-Peck, 2006; Lubienski, Lubienski & Crane, 2008; Sun, Bradley & Akers, 2012). The only study conducted in primary education that indicated a positive effect as well was the study by Borland & Howsen (2003). These authors also examined the curvilinear relationship of school size effects on academic achievement. The results of the two-stage least-squares regression suggested an optimal school size of around 760 students, which appeared to be much larger than the mean school size of 490 students found in the study.

Curvilinear Relationships (School Size as a Quadratic Function)

Besides Borland & Howsen, seven samples (3 studies) reported non-linear relationships as well (Bradley & Taylor, 1998; Foreman-Peck & Foreman-Peck, 2006; Sawkins, 2002). These studies are all conducted in secondary education in the United Kingdom, and all focused on the upper end of the exam results distribution. The results for the samples in England (Bradley & Taylor) and Wales (Foreman-Peck & Foreman-Peck) suggested an inverted `U' shaped relationship between school examination performance and school size, with optima around 1200 to 1500 students for schools in England and around 600 students for schools in Wales. In the study using Scottish data (Sawkins, 2002), a `U' shaped relationship was found. Scottish school examination performance appeared to decline as the number of pupils in a school increases, reaching a minimum turning point of around 1200 pupils, after which the performance started to increase. However, very large Scottish schools were uncommon. In the study by Sawkins only 4 per cent of the secondary schools appeared to be larger than the calculated minimum.

School Size Measured as Categories

In 15 studies (18 samples) schools were classified in categories, based on the numbers of pupils. Six studies (6 samples) were conducted in primary education and 10 studies (8 samples) in secondary education. The range of school sizes included in the studies was variable. Some studies compared small and larger schools while in other studies schools of three or more different size categories were compared.

The results of the vote count were mixed. In three samples (2 studies) a positive relationship between school size and achievement was found (large schools doing better) (Gardner, Ritblatt & Beatty, 2000; McMillen, 2004) and in three other samples (2 studies) a negative association (Eberts, Schwartz & Stone, 1990; Lee & Loeb, 2000) was established. In the majority of samples (16 samples) the relationship appeared to be non-significant. In the remaining six samples a certain size category or optimum was favored (Alspaugh, 2004; Lee & Smith, 1997; Ready & Lee, 2007; Rumberger & Palardy, 2005). For secondary education the size category most favored appeared to be mid-sized schools. The only study (sample) conducted in primary schools (Alspaugh, 2004) produced inconclusive results with only schools in the smallest size category (< 200 pupils) positively and significantly associated with achievement.

The study by Rumberger & Palardy (2005) needs further attention as it is one of the few studies that investigated the effects of school size on several outcome measures of high school performance (i.e. achievement growth, drop-out and transfer rate). The authors used data from the National Education Longitudinal Study (Nels:88) and applied multilevel analysis. The results showed that schools effective in promoting student learning (growth in achievement) not necessarily are effective in reducing drop-out and transfer rates as well. Achievement growth appeared to be significantly higher in large high schools (1200-1800 pupils) as was also the drop-out rate. Next to this, it was found that background characteristics contributed differently to the variability in the outcome measures (i.e. 58 per cent of the variance in school drop-out rates, 36 per cent of the variance in student achievement and 3 per cent of the variance in transfer) as did also school policies and practices. When dropout was the dependent variable, school policies and practices accounted for 25 per cent of the remaining variance after controlling for student background. This was far more than for achievement or transfer.

Moderator Analyses

For the studies and samples in which school size was measured as a continuous variable "moderator analyses" were conducted to examine the degree to which the relationship between school size and achievement would appear to be modified according to specific characteristics of the study or sample. It was also investigated whether the school size and achievement correlation was moderated by the academic subjects in the achievement measure.

Table 2.3

Results of vote counts examining the number and percentage of negative, non-significant and positive effects of school size on academic achievement in all subjects, language, mathematics, science and subjects other than math or language (school size measured as a continuous variable)

Subject	Negative effects N (%)	Non-significant effects N (%)	Positive effects N (%)
All subjects	20 (22%)	62 (69%)	8 (9%)
Subject Math	5 (20%)	19 (76%)	1 (4%)
Subject Language	7 (26%)	19 (74%)	0 (0%)
Subject Science	1 (17%)	4 (67%)	1 (17%)
Subject other than Math, Language or Science	7 (21%)	20 (61%)	6 (18%)

The results do not show differences of importance (see Table 2.3). The percentage of positive effects (students in larger schools having better performance) for achievement in "all other subjects" is somewhat higher, compared to those for mathematics.

Analyses of study and sample characteristics examining the number and percentage of negative, non-significant and positive effects of school size on academic achievement are presented in Table 2.4. The display of study and sample characteristics, the statistical technique employed and the inclusion of a covariate for student's prior achievement in the model tested show the most interesting variations.

Table 2.4

Results of “moderator analyses” examining the number and percentage of negative, non-significant and positive effects of school size on academic achievement (school size measured as continuous variable), for different study and sample characteristics

	Negative effects	Non-significant effects	Positive effects
“Moderator”	N (%)	N (%)	N (%)
Level of schooling			
Primary school	7 (22%)	24 (75%)	1 (3%)
Primary and secondary school	2 (40%)	3 (60%)	0 (0%)
Secondary school	11 (21%)	35 (66%)	7 (13%)
Country			
Canada	0 (0%)	1 (100%)	0 (0%)
Hong Kong	0 (0%)	0 (0%)	1 (100%)
Netherlands	0 (0%)	2 (100%)	0 (0%)
Sweden	0 (0%)	1 (100%)	0 (0%)
UK	2 (17%)	5 (42%)	5 (42%)
USA	18 (25%)	53 (73%)	2 (3%)
Covariates included			
Included covariate for student’s prior achievement	8 (33%)	15 (63%)	1 4(%)
Included covariate for ability	0 (0%)	3 (75%)	1 (25%)
Included covariate for SES	8 (24%)	23 (68%)	3 (9%)
Included covariate for composite SES	19 (23%)	57 (68%)	8 (11%)
Included covariate for urbanicity	2 (25%)	5 (63%)	1 (13%)
Statistical technique used			
Technique multilevel	7 (32%)	13 (59%)	2 (9%)
Technique not multilevel	13 (19%)	49 (72%)	6 (9%)
Total	20 (22%)	62 (69%)	8 (9%)

Relatively more negative effects are found in studies that account for prior achievement as well as in studies that employed multilevel modeling. The percentage of positive relationships found seems to be somewhat higher in secondary education compared to primary education. However, both at primary and secondary education level the analyses of study and sample characteristics suggests a negative tendency with relatively more studies yielding negative than positive effects.

Social Cohesion: Attitudes of Students and Teachers towards School

Fourteen studies (15 samples, yielding in total 26 effect estimates) provided evidence about the relationship between school size and students' and teacher attitudes towards school (see Table 2.6 and Table A2). Evidence about the effects of school size on attitudes was mainly available from secondary education (12 studies; 13 samples). Only two of the 14 studies examined the impact of school size on students' attitudes in primary education. Again most of the studies were conducted in the United States (9 studies; 10 samples). Other countries were Australia (1 study), Israel (1 study), Italy (1 study) and the Netherlands (2 studies).

The outcome variables (attitudes) measured in the studies could be classified into three main variables: identification and connection to school, relationships with students and relationships with teachers (see Table 2.5). With regard to students' identification and connectedness to schools the variables used included perceptions of pupils, like feeling part of the school, feeling competent and motivated, feeling safe, being happy and satisfied with school, with education and the usefulness of their school work in later life. Relationships with students were defined as perceptions of being happy together as well as the kindness and helpfulness of their peers. The relationship with teachers is a variable in which relational aspects were included (e.g. the teacher treats pupils fairly and cares about them) as well as perceptions with regard to the support students receive (such as encouraging students to higher academic performance, helping pupils with school work).

As identification and connection to school is concerned, Kirkpatrick Johnson et al. (2001) distinguish between affective aspects (the feelings towards and identification with school, which he calls school attachment) and behavioral aspects (students' participation or engagement). These authors refer to behaviors that represent participation, such as trying their best in class, doing homework, and participation in extra-curricular activities. In this section, where the attitudes of students and teachers towards school are the outcome variables, we limit ourselves to attitudes (or attachment) to identification of and connection with school. The effects of school size on participation will be discussed in a next section.

Table 2.5

Overview of outcome variables and variable heading used in studies where attitudes of students and teachers towards school were the dependent variable

	Variable	Variable heading
Student attitudes	Identification and connectedness to schools	School satisfaction (Bowen, Bowen & Richman, 2000)
		Student school attachment (Crosnoe, Kirkpatrick Johnson & Elder, 2004; Holas & Huston, 2012; Kirkpatrick Johnson, Crosnoe & Elder, 2001)
		Sense of belonging (Kahne, Sporte, De La Torre & Easton, 2008)
		Achievement motivation (Koth, Bradshaw & Leaf, 2008)
		School connectedness (McNeely, Nonnemaker & Blum, 2002; Van der Vegt, Blanken & Hoogeveen, 2005)
		Student engagement (Silins & Mulford, 2004)
		Students sense of community in the school (Vieno, Perkins, Smith & Santinello, 2005)
		Classroom climate (De Winter, 2003)
	Relationship with peers	Student engagement (Silins & Mulford, 2004)
		Students sense of community in the school (Vieno et al., 2005)
		Relationships with peers (Van der Vegt et al., 2005)
	Relationship with teachers	Teacher support (Bowen et al., 2000)
		Student-teacher bonding (Crosnoe et al., 2004)
		Student school attachment (Holas & Huston, 2012)
		Academic personalism, classroom personalism, student-teacher trust (Kahne et al., 2008)
		School connectedness (McNeely et al., 2002)
		Student engagement (Silins & Mulford, 2004)
		Students' sense of community in the school (Vieno et al., 2005)
		Relationships with teachers (Van der Vegt et al., 2005)
Teacher attitudes	Identification and connectedness to schools	Teachers' collective responsibility (Lee & Loeb, 2000)
		Communal school organization (Payne, 2012)
		Organizational commitment (Rosenblatt, 2001)
	Relationship with teachers	Teacher-teacher trust (Kanhne et al., 2008)
		Communal school organization (Payne, 2012)

Table 2.6 gives an overview of the number of studies, samples and estimates included in the vote-counting procedure for students' and teachers' attitudes to school. In total 14 studies and 14 samples were included in the vote count. Two-third of the effects (derived from half of the 14 samples) between school size and attitudes to school appeared to be negative, favoring small schools. None of the studies yielded positive effects.

Table 2.6
Results of vote counts examining the number of negative, non-significant, curvilinear and positive effects of school size on students' and teachers' attitudes to school

			Direction of effect			
	Studies	Samples	-	Ns	∩	+
School size measured as a continuous variable	11	12	12	7	0	0
School size measured as a quadratic function	1	1	0	0	1	0
School size measured as discrete variable (categories)	3	3	3	0	1	0
Total	14	15	15	7	2	0

- = negatively related with school size
ns = no significant relation with school size
∩ = optimal school size found
+ = positively related with school size

School Size Measured as a Continuous Variable
Nine studies reported linear effects of school size on attitudes to school. Five of these studies were conducted in the US, the other four in Australia, Israel, Italy and the Netherlands.

Mixed (both negative and not significant) effects were found in the studies by Crosnoe et al. (2004), Kahne et al. (2008) and Van der Vegt et al. (2005). Vieno et al. (2005) found a positive effect, although this effect was not significant. In the remaining five studies school size appeared to be (slightly) negatively associated with students' and teachers' attitudes.

One of the US studies in which a (small) negative effect was found is the study by McNeely et al. (2004). The authors used evidence from a sample taken from the National Longitudinal Study of Adolescent Health (about 75000 adolescents from 127 schools, grades 7-12). School size appeared to be negatively associated with school connectedness, but the strength of the relationship was meager, as an increase of 500 students was associated with a very small decline in school connectedness.

The study by Silins & Mulford (2004) was conducted in Australia. The authors applied path modeling to examine the association between school size and SES on both students' perceptions of teachers' work in the class and students' outcomes (such as attendance, participation in and engagement with school). Engagement with school was operationalized as students' perceptions with regard to the way teachers and peers relate to them, the usefulness of their schoolwork in later life, and the extent of identification with their school. School size had an indirect and significant negative effect on engagement through participation (i.e. absences, participation in extracurricular activities, preparedness to do extra school work, involvement in classroom decisions etc.). Students in large schools participated less and this was associated with less engagement.

In the study conducted in the Netherlands, mixed effects were found. Van der Vegt et al. (2005) reported a non-significant effect of school size on students' connectedness with school and significant negative effects of school size on both relationships with peers and relationships with teachers.

Curvilinear Relationships

The only study examining curvilinear relationships of students' and teachers' attitudes was the study by Crosnoe et al. (2004). The authors used data from the National Longitudinal Study of Adolescent Health. The sample included 15000 students from 84 secondary schools. The mean school size was 1381. Interpersonal climate was the dependent variable. It was measured with three variables, i.e. student school attachment, student-teacher bonding, and student extra-curricular participation. Multilevel modeling was applied to estimate the effects of school size. The amount of variation between schools appeared to be smaller for school attachment and teacher bonding (3 and 5 per cent respectively) than for extra-curricular participation (14 per cent). For school attachment and teacher bonding a curvilinear effect was found with the lowest levels of attachment and teacher binding occurring at a size of 1900 or 1700 students respectively. For extracurricular participation, a negative linear effect was found. The authors concluded that an optimal secondary school size for school connectedness would be less than 300 students, which is considerably lower than the optimal size for academic achievement found in other studies.

School Size Measured in Categories

In three of the four studies in which school size was measured in categories (Bowen et al., 2000; Lee & Loeb, 2000; Weiss, Carolan & Baker-Smith, 2010) significant negative associations were found in which small schools were favored over larger schools. The fourth study conducted by De Winter (2003) and employed in Dutch secondary education, favored mid-size schools. In this study it was concluded that, as far as school climate for pupils is concerned, a school should neither be too big nor too small.

Participation

Participation of students, teachers or parents was the dependent variable in 10 studies (see Table 2.7 and Table A3). With the exception of the study by Holas and Huston, in which primary and middle schools were sampled both, all studies were concerned with secondary education. Nine studies were conducted in the United States and one in Australia (Silins & Mulford, 2004).

Seven of the ten studies provided evidence on participation of students, one about participation of teachers and two about participation of parents (see Table 2.7). In five studies students' participation was restricted to participation in extracurricular activities; in the two remaining studies a broader operationalization of participation was taken. In the study by Holas and Huston school involvement included four aspects (school attachment, teacher support, negative affect towards school and school activity participation). Higher scores represented higher involvement. Silins and Mulford used a broad concept of students' participation as well which included absences, participation in extracurricular activities, preparedness to do extra work, involvement in classroom/school decisions and setting own learning goals, and voicing opinion in class.

The study by Kahne et al. (2008) examined the impact of four years of small school reform in Chicago. A variety of teacher and student measures was included in the study, including teachers' involvement in school decision making (see also the section on other dependent variables).

The impact of school size on participation of parents was examined in two studies. Dee, Ha & Jacob (2007) included four dependent variables about parental involvement in their study, each variable measured through one single item. The item addressing the most intense involvement with school (i.e. volunteering at school) was chosen to be included in this review.

Table 2.7
Overview of outcome variables and variable heading used in studies in which participation of students, teachers or parents was the dependent variable

	Variable	Variable heading
Participation of students	Extracurricular participation	Extracurricular participation (Coladarci & Cobb, 1996; Crosnoe et al., 2004; Feldman & Matjasko, 2006; Lay, 2007; McNeal, 1999)
	Broader school participation	School involvement including school activity participation (Holas & Huston, 2012)
		Participation in school activities (Silins & Mulford, 2004)
Participation of teachers	Involvement in school decision making	Teacher influence (Kahne et al., 2008)
Participation of parents		Parent(s) act as a volunteer at the school (Dee et al., 2007)
		Average of total number of California Parent Teacher Association members for each affiliated school (Gardner et al., 2000)

The results of the vote count for school size on participation are presented in Table 2.8.

In almost all samples a negative and significant association between size and participation was found despite different conceptualizations, outcome measurements and types of respondents. Although the number of studies is limited, such a pattern of results supports the claim that smaller schools are associated with greater engagement. This was also found in other review studies (see Leithwood et al., 2009).

Table 2.8

Results of vote counts examining the number of negative, non-significant, curvilinear and positive effects of school size on participation

			Direction of effect			
	Studies	Samples	-	ns	∩	+
School size measured as a continuous variable	7	8	8	0	0	0
School size measured as discrete variable (categories)	4	5	2	2	1	0
Total	10	10	10	2	1	0

-	= negatively related with school size
ns	= no significant relation with school size
∩	= optimal school size found
+	= positively related with school size

School Safety

Evidence about the relationship between school size and school safety was obtained from 24 studies (25 samples) (see Table 2.10 and Table A4). Two studies were conducted in primary education (Bonnet, Gooss, Willemen & Schuengel, 2009; Bowes, Arseneault, Maughan, Taylor, Caspi & Moffitt, 2009), one study used samples both from primary and secondary school students (O'Moore, Kirkham & Smith, 1997) and in three studies elementary and secondary school students were sampled together. The remaining 18 studies were conducted in secondary education. Thirteen studies were performed in the United States, five studies in the Netherlands (Bonnet et al., 2009; Inspectorate of Education, 2009; Mooij, Smeets & De Wit 2011; Van der Vegt et al., 2005; De Winter, 2003), two in Israel (Attar-Schwartz, 2009; Khoury-Kassabri, Benbenishty, Astor & Zeira, 2004), one in Ireland (O'Moore et al., 1997), one in the United Kingdom (Bowes et al., 2009), one in Canada (Leung & Ferris, 2008) and one in Taiwan (Wei, Williams, Chen & Chang, 2010).

The outcome variables addressed in the 24 studies referred to various forms of student safety behavior, including (combinations of) disciplinary behavior, bullying, norm violating behavior and different types of violence (see Table 2.9).

Table 2.9
Overview of outcome variables and variable heading used in studies in which safety was the dependent variable

Variable	Variable headings	Author(s)
Disciplinary school and class climate	School climate, respectful classroom behavior	Inspectorate of Education (2003); Kahne et al. (2008); Koth et al. (2008)
	Feelings of safety	Mooij et al. (2011)
	Students' behaviors (fights, use of alcohol, students' physical and verbal abuse of teachers etc.)	Bowen et al. (2009); Haller (1992)
	Misbehavior (disorder and bullying)	Chen (2008)
	School misbehavior	Stewart (2003)
Bullying	Bullying others and being bullies	Bowes et al. (2009); Klein & Cornell (2010); O'Moore et al. (1997); Van der Vegt et al. (2005); Wei et al. (2010); De Winter (2003)
Problem behavior	Norm violating behaviors, alcohol and marijuana	Chen & Vazsonyi (2013); Van der Vegt et al. (2005)
	Substance abuse while at school	Eccles, Lord & Midgely (1991)
	Suspensions	Heck (1993)
Violence	Sexual harassment	Attar-Schwartz (2009)
	Violence	Eccles et al. (1991); Leung & Ferris (2008); Van der Vegt et al. (2005); Watt (2003)
	Victimization (personal, property, physical, verbal)	Bonnet et al. (2009); Gottfredson & DiPietro (2011); Khoury-Kassabri et al. (2004); Klein & Cornell (2010)
	Crime (incidents)	Chen (2008); Chen & Weikart (2008)

The summary of directions of effect for school size and safety is presented in Table 2.10. The results indicate that the number of negative and the number of non-significant effects are about the same.

Table 2.10
Results of vote counts examining the number of negative, non-significant, curvilinear and positive effects of school size on safety)

			Direction of effect			
	Studies	Samples	-	ns	∩	+
School size measured as a continuous variable	17	17	19	17	0	5
School size measured as discrete variable (categories)	8	9	3	5	2	3
Total	24	25	21	22	2	9

-	= negatively related with school size
ns	= no significant relation with school size
∩	= optimal school size found
+	= positively related with school size

Positive Relationships/Mixed Effects
Positive effects of school size on feelings of safety were reported in five studies. With the exception of the study by O'Moore et al. (1997) in which a sample from primary and secondary schools was taken, all studies were conducted in secondary schools (De Winter, 2003; Klein & Cornell, 2010; Mooij et al., 2011; O'Moore, 1997; Van der Vegt, 2005).

Mooij et al. (2011) used data from almost 80,000 pupils, 6000 teachers and other staff and 600 managers from secondary school in the Netherlands to test a two level model of social cohesion influences on a pupil's feelings of school safety. The authors found a positive effect of school size: pupils felt more safely at larger schools. However, when interaction effects were added to the model (i.e. the interaction of school size with pupil social violence), the main effect for school size on pupil's feelings of safety became insignificant.

In another Dutch study, De Winter (2003 found positive effects as well. In this study being bullied, bullying and fighting occurred significantly more in smaller secondary schools, the same result was found after correction for level of attainment (school type, i.e. different streams of secondary education) or urbanicity. An explanation the author offered was that, as students in smaller schools do have more intense relationships with their peers, more frequent bullying and fighting obviously might also be part of these contacts.

The study by Klein and Cornell (2010) is the only one of the 13 US studies that also found positive effects for school size. In this study three types of victimization were the dependent variables (i.e. bullying, threats and physical attacks). Regression analysis was applied. The results were mixed, depending on the measurement of the outcome variable. When teacher and student perceptions of victimization were the dependent variable, the results indicated a negative effect (with significant higher levels of violence perceived in larger schools). Non-significant effects were found when student self-reports of being a victim of violence were used. And if violence rates based on school discipline records were the outcome measure, the results indicated a positive association. The contradictory findings suggest the need for a closer examination of the measures of victimization used.

Negative Relationships

A negative relation between school size and safety was reported in 11 studies (Attar-Schwartz, 2009; Bowen et al., 2000; Chen, 2008; Chen & Vazsonyi, 2013; Eccles et al., 1991; Leung & Ferris, 2008; Stewart, 2003; see also Bowes et al., 2009; Gottfredson & DiPietro, 2011; Haller, 1992; Van der Vegt et al., 2005).The effect might be small (with an increase of e.g. 500 pupils in a school increasing the risk for being a victim of bullying after controlling for neighborhood and family background variables and children's internalizing and externalizing behaviors, see e.g. Bowes et al., 2009), or partial, i.e. school size only matters for schools of a certain size category (see e.g. Leung & Ferris, 2008, in which only for very large schools a negative effect was found).

Leung & Ferris (2008) examined the effect of school size on self-reported teenage incidence of violence of 17 year old low SES French speaking males in Montreal, Canada, controlling for social and demographic characteristics. School size was measured both as a continuous variable and categorically, classified into four size categories. Depending on the measure of school size used, the results of the regression analysis differed. School size measured continuously was significantly (negatively) associated with teenage violence. When school size was measured discretely (broken down into four size categories) only for very large schools a negative effect was indicated. No significant effects were found for small and large medium sized schools.

School delinquency/misbehavior was the dependent variable in the study conducted by Stewart (2003). In this study school misbehavior was measured by means of a scale asking pupils how often during the first half of the current school year they got in trouble for not following school rules, were put on an in-school suspension, suspended or put on probation from school, or got into a physical fight at school. Multilevel modeling was applied to examine the effects six of school level and fourteen pupil level covariates on school misbehavior. Two school level variables in the model were significant: school size and school location. Larger schools in urban areas had significantly higher levels of school misbehavior. Higher levels of school attachment, school commitment and in especially beliefs in school rules appeared to be positively associated with lower levels of misbehavior as well.

Finally, Chen (2008) applied structural equation modeling to investigate how school size, school climate, and zero tolerance policies interact to affect school criminal incidents. The results showed school size to be positively associated with higher levels of school crime. School size also had indirect effects on school crime through school culture, which was operationalized in this study by discipline problems (misbehavior) and transience. According to Chen "reducing school size by and of itself may not prove to be effective in solving the school safety problems" (p. 315). Instead Chen recommends school reformers to "create opportunities for individual attention and student participation, which then lead to positive bonding and social culture, which in turn improve student behavior and reduce school crime" (p. 315).

Analyses of Study and Sample Characteristics

For the studies and samples in which school size was measured as a continuous variable "moderator analyses" were conducted to examine study and sample characteristics that may account for the differences of directions of school size effects (see Table 2.11).

Table 2.11

Results of analyses of study and sample characteristics examining the number and percentage of negative, non-significant and positive effects of school size on safety

	Negative effects	Non-significant effects	Positive effects
"Moderator"	N (%)	N (%)	N (%)
Level of schooling			
Primary school	1 (33%)	2 (66%)	0 (0%)
Primary and secondary school	3 (0%)	0 (100%)	0 (0%)
Secondary school	15 (44%)	14 (41%)	5 (15%)
Country			
Canada	1 (100%)	0 (0%)	0 (0%)
Israel	1 (20%)	4 (80%)	0 (0%)
Netherlands	2 (40%)	1 (20%)	2 (40%)
Taiwan	0 (0%)	2 (100%)	0 (0%)
UK	1 (33%)	2 (67%)	0 (0%)
USA	14 (54%)	8 (33%)	3 (13%)
Covariates included			
Included covariate for SES	9 (36%)	12 (48%)	4 (16%)
Included covariate for composite SES	14 (45%)	14 (45%)	3 (10%)
Included covariate for urban city	8 (53%)	3 (20%)	4 (27%)
Statistical technique used			
Technique multilevel	3 (23%)	9 (69%)	1 (8%)
Technique not multilevel	16 (57%)	8 (29%)	4 (14%)
Total	19 (46%)	17 (42%)	5 (12%)

The statistical technique employed and if a study was conducted in the United States are the most prominent outcomes. Relatively more negative effects are found in studies applied in the United States, as well as in studies that did not apply multilevel modeling. More significant effects (both negative and positive) were found if urbanicity was controlled for.

Student Absence and Dropout

Twelve studies (19 samples) reported on evidence about attendance, truancy or absenteeism. The effect of school size on dropout was examined in four studies (5 samples). Almost all studies (and samples) were conducted in secondary schools, with one study reporting evidence from primary schools (Durán-Narucki, 2008) and two studies employed in samples of both primary and secondary students (Eccles et al., 1991; Heck, 1993). With the exception of the study by Bos, Ruijters & Visscher (1990), conducted in the Netherlands and the study by Foreman-Peck and Foreman-Peck (2006) conducted in Wales (United Kingdom), all studies were conducted in the United States. Two studies (Gardner et al., 2000; Kahne et al., 2008) investigated the effect of size on both absenteeism and dropout.

The predominant outcome variables included in the studies referred to attendance and drop-out (see Tables 2.12 and 2.13, as well as Tables A5 and A6). Perceptions with regard to truancy and absenteeism were measured in just a few studies.

Table 2.12

Overview of outcome variables and variable heading used in studies in which attendance/absenteeism and truancy are the dependent variable

Variable	Variable headings	Author(s)
Truancy	Percentage of pupils absent	Bos et al., 1990
	Perceptions with regard to truancy	Haller, 1992
Attendance	Attendance rate	Chen & Weikart, 2008; Duran-Narucki, 2008; Foreman-Peck & Foreman-Peck, 2006; Heck, 1993; Jones, Toma & Zimmer, 2008; Kuziemko, 2006: Lee, Özgün-Koca & Cristol, 2011
Absenteeism	Absenteeism rate	Gardner et al., 2000; Kahne et al., 2008
	Perceptions with regard to absenteeism	Eccles et al., 1991

Table 2.13

Overview of outcome variables and variable heading used in studies in which dropout is the dependent variable

Variable	Variable headings	Author(s)
Drop-out	Drop-out rate	Gardner et al., 2000; Kahne, 2008; Lee & Burkam, 2003; Rumberger &Palardy, 2005

Before calculating the vote counts, the results of studies were rescored if necessary, so that in all cases a positive effect denotes a situation of high attendance and less absenteeism, truancy or drop-out.

Table 2.14 shows the summary of the vote counts for studies in which attendance or truancy were the dependent variable. One study (Durán-Narucki, 2008) reported a positive relationship between school size and attendance rate. Five studies reported negative effects (less attendance or absenteeism in larger schools) (Eccles et al., 1991; Foreman-Peck & Foreman-Peck, 2006; Gardner et al., 2000; Haller, 1992; Heck, 1993; Jones et al., 2008). Mixed effects were reported in three studies (Kahne et al., 2008, Kuziemko, 2006; Lee et al., 2011) and non-significant relationships in two studies as well (Bos et al., 1990; Chen & Weikart, 2008).

Table 2.14
Results of vote counts examining the number of negative, non-significant, curvilinear and positive effects of school size on attendance/absenteeism and truancy

			Direction of effect			
	Studies	Samples	-	ns	∩	+
School size measured as a continuous variable	11	18	10	10	0	2
School size measured as discrete variable (categories)	1	1	1	0	0	0
Total	12	19	11	10	0	2

-	= negatively related with school size
ns	= no significant relation with school size
∩	= optimal school size found
+	= positively related with school size

With regard to drop-out, three of the five studies reported significant differences between size categories. In the fourth study (Kahne et al., 2008), in which a linear effect of size was investigated, no statistically significant relationships were found (see also Table 2.15).

Table 2.15
Results of vote counts examining the number of negative, non-significant, curvilinear and positive effects of school size on drop-out

			Direction of effect			
	Studies	Samples	-	ns	∩	+
School size measured as a continuous variable	1	2	0	2	0	0
School size measured as discrete variable (categories)	3	3	1	0	2	0
Total	4	5	1	2	2	0

-	= negatively related with school size
ns	= no significant relation with school size
∩	= optimal school size found
+	= positively related with school size

Negative Relationships

Eccles et al. (1991) used data from the National Educational Longitudinal Study (NELS:88). They found absenteeism, violence and substance abuse significantly more often being reported as a major problem in larger schools by both teachers and students. Haller (1992) came to the same conclusion in his study into the effects of high school consolidations in rural areas on school level student indiscipline (truancy and vandalism/theft). The results showed that school size had a substantial effect on student indiscipline. However, testing the effect of school size in a cross section of rural schools the results showed that doubling the size of rural high schools did not affect school discipline in a substantial way. Therefore the author concluded that the decision underlying consolidations of rural high schools probably should rest on other criteria than its effect on student indiscipline.

Non-Significant Relationships

Chen & Weikart (2008) investigated the relationship between school size, school disorder, student attendance and achievement. The model builds upon the School Disorder Model by Welsh, Stokes and Greene (2000) and was extended for this study with student achievement. Participating schools were 212 middle schools in New York. Percentage free lunch and percentage white students were the control variables. Structural Equation Modeling was applied. Higher school disorder, a lower attendance rate and lower achievement were found in larger schools but the effects were not statistically significant. The hypothesis that "school size has an indirect effect on academic achievement mediated by school disorder and student attendance rate" could not be confirmed (p. 15). However, the results indicated a strong positive relationship between attendance rate and achievement. For policy implications, like Eccles et al., Chen & Weikart recommend to focus on measures to improve school climate, including attendance policies, instead of reducing school size.

School Size Measured as Categories

Three studies reported differences on attendance or dropout rate between various school size categories (Gardner et al., 2000; Lee & Burkam, 2003; Rumberger & Palardy, 2005). Gardner et al. compared small Californian public schools (between 200 and 600 pupils) and large schools (2000 pupils or more). Student achievement (four measures), absenteeism and dropout were the dependent variables. The results indicated a significant positive effect of school size on all student achievement measures. At the same negative effects were found for absenteeism and dropout. So students at larger schools performed better, but were more absent and dropout in large schools was significantly higher. This was also the conclusion in the study by Rumberger and Palardy (1995).

The study by Lee & Burkam (2003) built on the study by Rumberger (1995). Lee & Burkam also used the longitudinal data from the National Educational Longitudinal Study (NELS: 88). The sample consisted of 3840 students in 190 schools from the High School Effectiveness supplement of NELS:88. Whether or not a student dropped out between 10th and 12th grade was the dependent variable. Four categories of school size were compared (<600, 601-1500, 1501-2500, > 2500). The results indicated that compared to medium-sized schools (601-1500 pupils), large and very large schools have higher drop-out rates. Small schools also had higher dropout rates than medium-sized schools. Interaction effects indicated that in public or catholic schools of small and medium size with positive student-teacher relations, the probability on drop-out is less.

Other Student Outcome Variables

Six studies reported on school size effects on other student outcomes, i.e. student attitudes towards self and learning, and engagement (see Table 2.16 and Table A7). One of these studies collected data from primary schools and middle schools (Holas & Huston, 2012), the remaining studies all included evidence from secondary schools. One study (Inspectorate of Education, 2003) was conducted in the Netherlands, the other six studies in the United States.

Table 2.16

Overview of variables and variable heading used in studies on other student outcome variables

Variable	Variable headings	Author(s)
Attitudes	Pupil attitudes towards self or learning	Self-esteem (Coladarci & Cobb, 1996) Perceived efficacy and competence in English and math (Holas & Huston, 2012)
Behavior	Engagement	Engagement in school (Kirkpatrick Johnson et al, 2001); Academic engagement (Lee & Smith, 1995) Participation in community services (Lay, 2007) School engagement (Weiss, Carolan & Baker-Smith, 2010)

The results were mixed (see Table 2.17). Two studies (Coladarci & Cobb, 1996; Holas & Huston, 2012) reported non-significant relationships between school size and student outcomes, two other studies reported negative effects (Lay, 2007; Weiss et al., 2010). For one study (Kirkpatrick Johnson et al., 2001), a non-significant effect was found at the primary level, while at the secondary level larger schools were associated with less student engagement.

Table 2.17

Results of vote counts examining the number of negative, non-significant, curvilinear and positive effects of school size on other student outcome variables

			Direction of effect			
	Studies	Samples	-	ns	∩	+
School size measured as a continuous variable	4	5	2	3	0	0
School size measured as discrete variable (categories)	3	3	1	1	2	0
Total	5	7	3	4	2	0

Attitudes

Two studies, one in US middle and one in US high schools investigated the relationship between school size and student attitudes. Coladarci & Cobb (1996) examined the indirect effect of school size on 12th grade academic achievement and self-esteem through (total time spent on) extracurricular participation. Using evidence from the National Educational Longitudinal Study of 1988 database, only students that attended either a small high school (less than 800 pupils) or a large high school (1600 or more pupils) were considered in the study. Structural equation modelling was applied. Variables included in the model were prior self-esteem and prior achievement, SES, size, total extracurricular participation and total time spent on extracurricular participation. The authors did find a significant negative effect of school size on extracurricular participation, with higher extracurricular participation among students attending smaller schools. The indirect effects of school size on achievement and self-esteem through extracurricular participation were negative, but not significant.

Holas & Huston (2012) applied path analysis to compare student achievement, school engagement and perceived efficacy and competence in English and math of students starting middle schools in 5th and 6 grades compared to students of the same grade in elementary schools. School characteristics (observed classroom quality, teacher related classroom quality, school percentage of minority and poor students, and school size) were included in the path model as intermediate variables. The authors did not find significant effects of school size on any of the outcome variables of students in 5th grade. In 6th grade school size was negative and significantly related to school engagement.

Engagement

Three studies investigated the impact of school size on student engagement in schools (Kirkpatrick Johnson et al., 2001; Lee & Smith, 1995; Weiss et al., 2010). In these studies engagement in school was operationalized in very different ways. Lee & Smith (1995) used the concept academic engagement, a composite of eight items measuring student behavior related to work in class. Kirkpatrick Johnson et al. (2001) focused on engagement in school (operationalized as attendance, attention for school work and doing homework), while Weiss et al. (2010) used a very broad composite measure of engagement based on seven variables: teacher experience, delinquent behavior, academic friend, educational motivation, teachers' belief about ability, school preparedness and parental involvement.

Weiss et al. (2010) investigated the impact of size on achievement and engagement in US high schools. Using data from the Educational Longitudinal Study (ELS 2002) they found that there are significant differences related to student engagement between schools of different size categories, while school size is not significantly related to mathematics achievement. Compared with students attending schools of the smallest size (the omitted category in the multilevel analysis), students in mid-sized or large schools appeared to have (significant) lower levels of engagement.

School Organization and Teaching and Learning

Three studies in the review included measures of the impact of school size on school organization and teaching and learning (see Table 2.18 and Table A8). These studies had different aims and scope.

Table 2.18

Overview of outcome variables and variable heading used in studies on school organization and teaching and learning

Variable	Variable headings	Author(s)
Teaching and learning	Expectations and support	Expectations for postsecondary education, academic press, peer support for academic achievement, school-wide future orientation (Kahne et. al, 2008);
	Instruction	Pedagogical and didactical approach (Inspectorate of Education, 2003);
		Quality student discussions in classroom, quality English instruction, quality Math instruction (Kahne et. al, 2008);
		Teachers' work (Silins & Mulford, 2004)
School organization	Teacher attitudes	Teacher efficacy (Eccles et al., 1991)
		Teachers' collective responsibility, commitment to innovation (Kahne et. al, 2008)
	Leadership	Principal instructional leadership (Kahne et. al, 2008);
		Teacher Leadership (Silins & Mulford, 2004)
	Curriculum	Program coherence (Kahne et. al, 2008);
	Professional development	Quality professional development, reflective dialogue (Kahne et. al, 2008)
	Organizational learning	Organizational learning (Silins & Mulford, 2004)

Thirteen of the 17 effects reported are derived from the study by Kahne et al.(2008), three from the study of Silins and Mulford (2004), and two from the study by Eccles et al. (1991) and the study of the Dutch Inspectorate of Education (2003), respectively. The results of the vote counts are mixed: most effect sizes appeared to be not significant, six effects reported were negative (favoring small schools) and for one study a curvilinear relationship was found (see Table 2.19).

Table 2.19
Results of vote counts examining the number of negative, non-significant, curvilinear and positive effects of school size on school organization and teaching and learning

			Direction of effect			
	Studies	Samples	-	Ns	∩	+
School size measured as a continuous variable	3	3	6	11	0	0
School size measured as discrete variable (categories)	1	1	0	0	1	0
Total	4	4	6	11	1	0

-	= negatively related with school size
ns	= no significant relation with school size
∩	= optimal school size found
+	= positively related with school size

Negative and Non-Significant Relationships

The study by Kahne et al. (2008) focused on the implementation and impact of the first phase of the Chicago High School Redesign Initiative (CHSRI). A theoretical framework summarizing the theory of change underlies this study and portrays the mechanisms through which the characteristics of small school reform are thought to promote a supportive and personalized context for students as well as a desirable teacher context for reform, which in turn would impact on instruction and different types of student outcomes (absences, drop-out rate, graduation rate and achievement test scores). The results of the three level multilevel analysis yielded four significantly negative effects and nine non-significant effects. It was found that teachers in CHSRI schools had a better context for reform (significantly greater level of commitment to innovation and a higher sense of collective responsibility). CHSRI schools also provided a more supportive context for students (with significantly higher expectations for post-secondary education and school-wide future orientation, but no significant difference for peer support for academic achievement). However, after the first phase, the improved contexts for teacher and students in CHSRI schools did not have a statistically significant impact on facilitators for instructional improvement (principal leadership, professional development, program coherence) and improved instructional practices (quality of student discussions, quality of

English and math instruction, academic press). So although some significant positive indications of the effects Chicago High School Redesign Initiative were visible, after five years it still “might be too soon to make broad claims about the efficacy of small school conversions in Chicago” (p. 299).

Silins & Mulford (2004) employed path modeling to examine the impact of school external (size and SES) and school internal variables on teacher leadership, organizational learning, teachers’ work and ultimately students’ outcomes (i.e. participation in and engagement with school).The study was conducted in Australia. School size had a significant negative indirect effect on organizational learning through staff perceptions of the availability of resources. School size was not significantly associated with teacher leadership and teachers‘ work.

Curvilinear Relationship

The study of the Dutch Inspectorate of Education (2003) had the aim to investigate the associations between various aspects of the quality of Dutch secondary schools as assessed by the Inspectorate (such as achievement, pedagogical and didactical approach, pupil guidance and quality care) and elements of school structure (size, school types, locations). In this study a curvilinear relationship was found between school size and the quality of the pedagogical and didactical approach. The results indicated that midsize schools (500-1000 pupils) had the lowest score on the quality of the pedagogical and didactical approach.

Costs

The review on costs was limited to studies that investigated variations in per pupil expenditure between schools of different sizes. Studies in which costs were measured at the above school level (at the district level for example as in Chakraborty, Biswas & Lewis, 2000) were excluded.

Five studies investigated variations in economic outcomes at school level (see Table A9). Four studies were from the USA and one from the Netherlands. Two studies were conducted in primary education (Merkies, 2000; Stiefel, Berne, Iatarola & Fruchter, 2000), one in secondary education (Bickel, Howley, Williams & Glascock, 2001) and two studies relate to both primary and secondary education (Bowles & Bosworth, 2002; Lewis & Chakraborty, 1996).

All studies reported a significant negative effect of school size on costs per pupil (Bickel et al, 2001; Bowles & Bosworth, 2002; Lewis & Chakraborty, 1996; Merkies, 2000; Stiefel et al., 2000) (see Table 2.20). A similar pattern was reported in each study. Sharp decreases in per pupil expenditure occur as the school size increases from very low to average, whereas the increase from average onwards is associated with much more modest decreases in costs. All studies take into account the impact of student population characteristics (e.g. income and ethnicity) and educational output (e.g. achievement scores, dropout or graduation rates) when assessing the effect of school size on costs per student. The effect of school size remains intact when controlling for educational output. In the study by Stiefel et al. (2000), however, the effect of school size largely disappears when taking into account student population characteristics (especially limited English proficiency).

Table 2.20
Results of vote counts examining the number of negative, non-significant, curvilinear and positive effects of school size on costs

			Direction of effect			
	Studies	Samples	-	ns	∩	+
School size measured as a continuous variable	4	4	4	0	0	0
School size measured as discrete variable (categories)	1	1	0	1	0	0
Total	5	5	4	1	0	0

-	= negatively related with school size
ns	= no significant relation with school size
∩	= optimal school size found
+	= positively related with school size

Conclusion and Discussion

The overall pattern of the vote-counting procedure show that, across all studies that examined the association between school size and any dependent variables, almost half (49%) of the effect estimates appeared to be non-significant, and one third (34%) negative (see Table 2.21). Positive effect relationships and non-linear relationships were found for 8 per cent of each of these two estimates.

Based on these results we cannot conclude that smaller schools are generally better for all types of outcomes. For certain non-cognitive outcomes, i.e. social cohesion or participation of students or parents in school activities, the findings in the review are consistent and indeed clearly suggest a positive impact of smaller schools. For other non-cognitive outcomes, like safety and school attendance, however, the number of negative and non-significant findings did not differ that much from each other. Although the empirical evidence of school size on safety tends to be negative as well (with more safety in smaller schools), the results are less convincing than appears to be the case for attitudes and participation. For safety some positive effects were found as well (17% of the estimates, derived from five studies), while such positive associations (favoring larger schools) did not occur for attitudes of students and teachers and participation.

When it comes to academic outcomes, our results suggest that “size does not matter”. When student achievement was the outcome measure, two third of the reported school size effects failed to reach statistical significance, 18 per cent were negative, 9 per cent positive and for 11% of the effects a curvilinear effect was found. In secondary education for those studies that reported curvilinear effects the optimal school size found was between 1100 and 1400 students on average.

The association between school size and school organization and teaching and learning was investigated in three studies. The majority of effects reported (13 out of 17) are derived from one study. As for achievement the results are mixed, with more than half of the estimates being non-significant.

Table 2.21

Directions of effect of school size on various dependent variables

			Direction of effect			
	Studies	Samples	-	ns	∩	+
Dependent variable			N (%)	N (%)	N (%)	N (%)
Achievement	46	64	23 (18%)	78 (62%)	14 (11%)	11 (9%)
Students' and teachers' attitudes to school	14	15	15 (63%)	7 (29%)	2 (8%)	0 (0%)
Participation	10	10	10 (77%)	2 (15%)	1 (8%)	0 (0%)
Safety	24	25	21 (39%)	22 (41%)	2 (4%)	9 (17%)
Attendance/absenteeis m and truancy	12	19	11 (48%)	10 (43%)	0 (0%)	2 (9%)
Drop-out	4	5	1 (20%)	2 (40%)	2 (40%)	0 (0%)
Other student outcome variables (attitudes towards self and learning, engagement)	5	6	3 (33%)	4 (44%)	2 (22%)	0 (0%)
School organization and teaching and learning	4	4	6 (33%)	11 (61%)	1 (6%)	0 (0%)
Costs	5	5	4 (80%)	1 (20%)	0 (0%)	0 (0%)
Total	84	107	94 (34%)	137 (49%)	23 (8%)	23 (8%)

- = negatively related with school size
ns = no significant relation with school size
∩ = optimal school size found
+ = positively related with school size

For academic achievement and safety, results were disaggregated for study characteristics, for those studies and samples in which school size was measured as a continuous variable. This approach was seen as a surrogate for moderator analysis in quantitative meta-analysis. For academic achievement the most striking outcomes of these analyses concerned the statistical technique employed and the inclusion of a covariate for student's prior achievement in the model. Relatively more negative effects were found in studies, which accounted for prior achievement and in studies that employed multilevel modeling. For safety as the dependent variable the statistical technique used and the country in which a study was conducted are the most prominent outcomes of the moderator analysis. Relatively less negative and more insignificant findings were found when multilevel modeling was applied. Studies conducted in the United States yielded relative more negative effects as compared to studies employed in other countries.

The review of costs was limited to studies that investigated variations in per pupil expenditure between schools of different sizes. All five studies included in the review reported a negative effect of school size on costs per pupil, be it that in the study by Stiefel et al. (2000) the effect of school size became insignificant when controlling for student population characteristics. The pattern reported in each study was in the same direction: sharp decreases in per pupil expenditure occur as the school size increases from very low to average, whereas the increase from average onwards is associated with much more modest decreases in costs. This conclusion is based on studies that took only student achievement or student graduation into account as control variables. Smaller schools might be more efficient, possibly also due to lower drop-out rates.

The results of the vote count analyses that were reported in this chapter confirm the outcomes of a review of earlier meta-analyses (Scheerens, Hendriks & Luyten, 2014b) and recent quantitative analyses of school size effects (Luyten, 2014). The quantitative summary of school size effects reported by Luyten was based on a subset of the studies included in this chapter. For studies that provided sufficient quantitative information and that controlled either for previous achievement (if achievement was the outcome measure) or socio-economic background (in case of non-cognitive outcomes) Luyten calculated standardized outcomes for different school sizes. The results were presented separately for primary and secondary education. With respect to the impact of school size on cognitive outcomes Luyten reports a negative and very weak effect in primary education. For secondary education the effect was very small as well, but curvilinear, with the highest scores in schools with between 1200 and 1600 students. These findings roughly confirm the results found in this chapter. In our review, the majority of studies reported either non-significant effects or negative effects. When studies analyzed quadratic relationships or compared school size categories, the optimal secondary school size appeared to be mid-sized (within the range of 1100-1400 pupils). However, it should be noted that the number of studies in our review that established non-linear associations was limited. Our results also confirm the findings of previous review studies (Leithwood & Jantzi, 2009; Newmann et al., 2006). While Newman et al. only focused on secondary school size, Leithwood and Jantzi examined the impact of school size in primary education as well. In both reviews, the results were mixed, with in secondary education the most defensible solution favoring mid-size schools (optimal school size of 1000 pupils) and in primary education smaller schools (optimal school size of 500 pupils).

Where, in this chapter, a distinction was made between various types of non-cognitive outcomes Luyten (2014) did not differentiate among various types of non-cognitive outcomes, and just distinguished cognitive and non-cognitive outcomes. Taking all studies examining non-cognitive outcomes in primary education together, the effect of school size that Luyten reports was negative, and fairly weak, which is comparable to what we found for primary and secondary education together. For secondary education Luyten found a small effect as well but slightly in favor of larger schools. However, when the summary was based on American studies solely, a reverse trend became apparent. The latter corresponds to what we found in our "moderator analyses" on safety and what was found in previous meta-analyses as well, with relatively more negative effects reported in studies conducted in the United States (see Scheerens, Hendriks & Luyten, 2014b).

A limitation of the current review and previous reviews as well (see e.g. Leithwood & Jantzi, 2009) is that the vast majority of the empirical evidence about school size effects is derived from studies conducted in the United States. Fifty-eight of the 84 studies included in our review were conducted in the United States. For academic achievement this was even more the case as only eight of the 46 studies examining the impact of size on cognitive outcomes did not originate from the United States.

A further drawback, and this refers to the third research question for this review, is that we still know very little on the indirect effects of school size effects. Previous research suggests that preconditions and intermediating school organization, teaching and learning factors affect the path from school size to cognitive and non-cognitive outcomes. From previous reviews we know that school size matters more for disadvantaged than for average students. Results from previous reviews and this review also suggest that school size effects depend on age level, and differ for rural and urban contexts and between countries. However, we still have little knowledge on the causal mechanisms that account for the assumed relationship between school size and outcomes. Just a few studies in the current review applied research methods such as structural equation modeling to test plausible mechanisms through which significant preconditions or intermediate variables (such as school climate, attendance policies, extracurricular participation and organizational learning) interact to produce school size effects (see Chen, 2008; Chen & Weikart, 2008; Coladarci & Cobb, 1996; Silins & Mulford, 2004). Another empirical study addressing indirect effects but not included in the current review due to a different operationalization of school size (Opdenakker & Van Damme, 2007) found the positive effect of large schools mediated by better teacher cooperation and classroom climate. For further research, using multilevel, longitudinal or experimental data and extended indirect effect models, where school size effects are hypothesized as being mediated by conditions at school and classroom level, are seen as relevant to try and break open the black box of positive, negative, curvilinear and non-significant school size effects.

References

Andrews, M., Duncombe, W., & Yinger, J. (2002). Revisiting economies of size in American education: are we any closer to a consensus? *Economics of Education Review*, *21*, 245-262. doi:10.1016/S0272-7757(01)00006-1

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Bushman, B. J. (1994). Vote-counting procedures in meta-analysis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 193-213). New York: Russell Sage Foundation.

Bushman, B .J., & Wang, M. C. (2009). Vote-counting procedures in meta-analysis. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.) (pp. 208-222). New York: Russell Sage Foundation.

Chakraborty, K., Biswas, B., & Lewis, W. C. (2000). Economies of scale in public education: an econometric analysis. *Contemporary Economic Policy*, *18*, 238-247. doi:10.1111/j.1465 -7287.2000.tb00021.x

Cotton, K. (2001). *New small learning communities: Findings from recent literature* (Vol. 40). Northwest Regional Educational Laboratory Portland, OR.

Hendriks, M. A. (2014). Research synthesis of studies published between 1990 and 2012. In H. Luyten, M. A. Hendriks & J. Scheerens (Eds.), *School size effects revisited* (SpringerBriefs in Education) (pp. 41-175). Cham: Springer .

Hendriks, M. Scheerens, J., & Steen, R. (2008). *Schaalgrootte en de menselijke maat*. Enschede: Universiteit Twente.

Kahne, J. E., Sporte, S. E., De La Torre, M., & Easton, J. Q. (2008). Small high schools on a larger scale: The impact of school conversions in Chicago. *Educational Evaluation and Policy Analysis*, *30*, 281-315. doi:10.3102/0162373708319184

Leithwood, K., & Jantzi, D. (2009). A review of empirical evidence about school size effects: A policy perspective. *Review of Educational Research*, *79*, 464-490. doi:10.3102/0034654308326158

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Luyten, H. (2014). Quantitiative summary of research findings. In H. Luyten, M. A. Hendriks & J. Scheerens (Eds.), *School size effects revisited* (SpringerBriefs in Education) (pp. 177-218). Cham: Springer.

Newman, M., Garrett, Z., Elbourne, D., Bradley, S., Noden, P., Taylor, J., & West, A. (2006). Does secondary school size make a difference: A systematic review? *Educational Research Review*, *1*, 41-60. doi:10.1016/j.edurev.2006.03.001

NWO (2011). Programma voor Onderwijsonderzoek (PROO) – Review Studies. Call for proposals 2011. Den Haag: Nederlandse organisatie voor Wetenschappelijk Onderzoek.

Onderwijsraad (2005). *Variëteit in schaal. Keuzevrijheid, sociale samenhang en draagvlak bij grote organisaties*. Den Haag: Onderwijsraad.

Opdenakker, M. C., & Van Damme, J. (2007). Do school context, student composition and school leadership affect school practice and outcome in secondary education? *British Educational Research Journal*, *33*, 179-206. doi:10.1080/01411920701208233

Rumberger, R. W. (1995). Dropping out of middle school: A multilevel analysis of students and schools. *American Educational Research Journal*, *32*, 583–625. doi:10.3102

/00028312032003583

Scheerens, J., Hendriks, M. A., & Luyten, H. (2014). Introduction. In H. Luyten, M. A. Hendriks & J. Scheerens (Eds.), *School size effects revisited* (SpringerBriefs in Education) (pp. 1-5). Cham: Springer.

Scheerens, J., Hendriks, M. A., & Luyten, H. (2014). School size effects: review and conceptual analysis. In H. Luyten, M. A. Hendriks & J. Scheerens (Eds.), *School size effects revisited* (SpringerBriefs in Education) (pp. 7-39). Cham: Springer .

Scheerens, J., Luyten, H., Steen, R., & Luyten-de Thouars, Y. (2007). *Review and meta analyses of school and teaching effectiveness*. Enschede: Department of Educational Organisation and Management, University of Twente.

Scheerens, J., Seidel, T., Witziers, B., Hendriks, M., & Doornekamp G. (2005). *Positioning and validating the supervision framework.* Enschede: University of Twente, Department of Educational Organization and Management.

Welsh, W. N., Stokes, R., & Greene, J. R. (2000). A macro-level model of school disorder. *Journal of Research in Crime and Delinquency*, *37*, 243-283.

Studies Used for Vote-Count

Åberg-Bengtsson, L. (2004). Do small rural schools differ? A comparative two-level model of reading achievement among Swedish 9-year-olds. *Scandinavian Journal of Educational Research*, *48*, 19-33. doi:10.1080/0031383032000149823

Alspaugh, J. W. (2004). School size as a factor in elementary school achievement. *ERS spectrum*, *22*(2), 28-34.

Archibald, S. (2006). Narrowing in on educational resources that do affect student achievement. *Peabody Journal of Education*, *81*, 23-42. doi:10.1207 /s15327930pje8104_2

Attar-Schwartz, S. (2009). Peer sexual harassment victimization at school: The roles of student characteristics, cultural affiliation, and school factors. *American Journal of Orthopsychiatry*, *79*, 407-420. doi:10.1037/a0016553

Barnes, J., Belsky, J., Broomfield, K. A., & Melhuish, E. (2006). Neighbourhood deprivation, school disorder and academic achievement in primary schools in deprived communities in England. *International Journal of Behavioral Development*, *30*, 127-136. doi:10.1177/0165025406065385

Bickel, R., Howley, C., Williams, T., & Glascock, C. (2001). High school size, achievement equity, and cost: Robust interaction effects and tentative results. *Education Policy Analysis Archives*, *9*(40). Retrieved from http://epaa.asu.edu/ojs/article/view/369/495

Bonnet, M., Gooss, F. A., Willemen, A. M., & Schuengel, C. (2009). Peer victimization in Dutch school classes of four- to five-year-olds: Contributing factors at the school level. *The Elementary School Journal, 110*, 163-177. doi:10.1086/605769

Borland, M. V., & Howsen, R. M. (2003).An examination of the effect of elementary school size on student academic achievement. *International Review of Education*, *49*, 463-474.

Bos, K. T., Ruijters, A., & Visscher, A. (1990). Truancy, drop-out, class repeating and their relation with school characteristics. *Educational Research*, *32*, 175-185. doi:10.1080 /0013188900320302

Bowen, G. L., Bowen, N. K., & Richman, J. M. (2000). School size and middle school students' perceptions of the school environment. *Social Work in Education*, *22*, 69-82.

Bowes, L., Arseneault, L., Maughan, B., Taylor, A., Caspi, A., & Moffitt, T. E. (2009). School, neighborhood, and family factors are associated with children's bullying involvement: A nationally representative longitudinal study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *48*, 545-553.

Bowles, T. J., & Bosworth, R. (2002). Scale economies in public education: Evidence from school level data. *Journal of Education Finance*, *28*, 285-299.

Bradley, S., & Taylor, J. (1998). The effect of school size on exam performance in secondary schools. *Oxford Bulletin of Economics and Statistics*, *60*, 291-324. doi:10.1111/1468-0084.00102

Caldas, S. J. (1993). Reexamination of input and process factor effects on public school achievement. *The Journal of Educational Research*, *86,* 206-214. doi:10.1080 /00220671.1993.9941832

Carolan, B. V. (2012). An examination of the relationship among high school size, social capital, and adolescents' mathematics achievement. *Journal of Research on Adolescence*, *22*, 583-595. doi:10.1111/j.1532-7795.2012.00779.x

Chen, G. (2008). Communities, students, schools and school crime - A confirmatory study of crime in US high schools. *Urban Education*, *43,* 301-318. doi:10.1177 /0042085907311791

Chen, G., & Weikart, L. A. (2008). Student background, school climate, school disorder, and student achievement: An empirical study of New York city's middle schools. *Journal of School Violence*, *7*(4), 3-20. doi:10.1080/15388220801973813

Chen, P., & Vazsonyi, A. T. (2013). Future orientation, school contexts, and problem behaviors: A multilevel study. *Journal of Youth and Adolescence*, *42*, 67-81. doi:10.1007s10964-012-9785-4

Coladarci, T., & Cobb, C. D. (1996). Extracurricular participation, school size, and achievement and self-esteem among high school students: A national look. *Journal of Research in Rural Education*, *12*, 92-103.

Crosnoe, R., Kirkpatrick Johnson, M., & Elder, G. H. (2004). School size and the interpersonal side of education: An examination of race/ethnicity and organizational context. *Social Science Quarterly*, *85*, 1259-1274. doi:10.1111/j.0038-4941.2004.00275.x

Dee, T. S., Ha, W., & Jacob, B. A. (2007). The effects of school size on parental involvement and social capital: Evidence from the ELS:2002. In T. Loveless & F. Hess (Eds.), *Brookings papers on education Policy* (pp. 77-97). Washington DC: Brookings Institution Press.

Deller, S. C., & Rudnicki, E. (1993). Production efficiency in elementary education: The case of Maine public schools. *Economics of Education Review*, *12*, 45-57. doi:10.1016/0272

-7757(93)90042-F

Driscoll, D., Halcoussis, D., & Svorny, S. (2003). School district size and student performance. *Economics of Education Review*, *22*, 193-201. doi:10.1016/S0272-7757(02)00002-X

Durán-Narucki, V. (2008). School building condition, school attendance, and academic achievement in New York City public schools: A mediation model. *Journal of Environmental Psychology*, *28,* 278-286. doi:10.1016/j.jenvp.2008.02.008

Eberts, R. W., Schwartz, E. K., & Stone, J. A. (1990). School reform, school size, and student achievement. *Economic Review*, *26*(2), 2-15.

Eccles, J. S., Lord, S., & Midgely, C. (1991). What are we doing to early adolescents? The impact of educational contexts on early adolescents. *American Journal of Education*, *99*, 521–542.

Feldman, A. F., & Matjasko, J. L. (2007). Profiles and portfolios of adolescent school-based extracurricular activity participation. *Journal of Adolescence*, *30*, 313-332. doi:10.1016 /j.adolescence.2006.03.004

Fernandez, K. E. (2011). Evaluating school improvement plans and their affect on academic performance. *Educational Policy*, *25*, 338-367. doi:10.1177/0895904809351693

Foreman-Peck, J., & Foreman-Peck, L. (2006). Should schools be smaller? The size-performance relationship for Welsh schools. *Economics of Education Review*, *25*, 157-171. doi:10.1016/j.econedurev.2005.01.004

Fowler, W. J., & Walberg, H. J. (1991). School size, characteristics, and outcomes. *Educational Evaluation and Policy Analysis*, *13*, 189-202. doi:10.2307/1164583

Gardner, P. W., Ritblatt, S. N., & Beatty, J. R. (2000). Academic achievement and parental involvement as a function of high school size. *The High School Journal*, *83*(2), 21-27.

Gottfredson, D. C., & DiPietro, S. M. (2011). School size, social capital, and student victimization. *Sociology of Education*, *84*, 69-89. doi:10.1177/0038040710392718

Haller, E. J. (1992). High-school size and student indiscipline: Another aspect of the school consolidation issue. *Educational Evaluation and Policy Analysis*, *14*, 145-156. doi:10.3102/01623737014002145

Heck, R. H. (1993). School characteristics, school academic indicators and student outcomes: implications for policies to improve schools. *Journal of Education Policy*, *8*, 143-154. doi:10.1080/0268093930080203

Holas, I., & Huston, A. C. (2012). Are middle schools harmful? The role of transition timing, classroom quality and school characteristics. *Journal of Youth and Adolescence*, *41*, 333-345. doi:10.1007/s10964-011-9732-9

Inspectie van het Onderwijs (2003). *Schoolgrootte en kwaliteit. Groot in kleinschaligheid*. Utrecht: Inspectie van het Onderwijs.

Jones, J. T., Toma, E. F., & Zimmer, R. W. (2008). School attendance and district and school size. *Economics of Education Review*, *27*, 140-148. doi:10.1016/j.econedurev.2006. 09.005

Kahne, J. E., Sporte, S. E., De La Torre, M., & Easton, J. Q. (2008). Small high schools on a larger scale: The impact of school conversions in Chicago. *Educational Evaluation and*

Policy Analysis, *30*, 281-315. doi:10.3102/0162373708319184

Khoury-Kassabri, M., Benbenishty, R., Astor, R. A., & Zeira, A. (2004). The contributions of community, family, and school variables to student victimization. *American Journal of Community Psychology*, *34,* 187-204. doi:10.1007/s10464-004-7414-4

Kirkpatrick Johnson, M., Crosnoe, R., & Elder Jr, G. H. (2001). Students' attachment and academic engagement: the role of race and ethnicity. *Sociology of Education*, *74*, 318–340. doi:10.2307/2673138

Klein, J., & Cornell, D. (2010). Is the link between large high schools and student victimization an illusion? *Journal of Educational Psychology*, *102*, 933-946. doi:10.1037/a0019896

Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2008). A multilevel study of predictors of student perceptions of school climate: The effect of classroom-level factors. *Journal of Educational Psychology*, *100*, 96-104. doi:10.1037/0022-0663.100.1.96

Kuziemko, I. (2006). Using shocks to school enrollment to estimate the effect of school size on student achievement. *Economics of Education Review*, *25*, 63-75. doi:10.1016 /j.econedurev.2004.10.003

Lamdin, D. J. (1995). Testing for the effect of school size on student achievement within a school district. *Education Economics*, *3*, 33-42. doi:10.1080/09645299500000002

Lay, J. C. (2007). Smaller isn't always better: School size and school participation among young people. *Social Science Quarterly*, *88*, 790-815. doi:10.1111/j.1540-6237.2007 .00483.x

Lee, V. E., & Burkam, D. T. (2003). Dropping out of high school: The role of school organization and structure. *American Educational Research Journal*, *40*, 353-393. doi:10.3102/00028312040002353

Lee, V. E., & Loeb, S. (2000). School size in Chicago elementary schools: Effects on teachers' attitudes and students' achievement. *American Educational Research Journal*, *37*, 3-31. doi:10.3102/00028312037001003

Lee, H. J., Özgün-Koca, S. A., & Cristol, D. (2011).An analysis of high school transformation effort from an outcome perspective. *Current Issues in Education*, *14*, 1-33. Retrieved from http://cie.asu.edu/ojs/index.php/cieatasu/article/view/

Lee, V. E., & Smith, J. B. (1995). Effects of high school restructuring and size on early gains in achievement and engagement. *Sociology of Education*, *68*, 241-270. doi:10.2307 /2112741

Lee, V. E., & Smith, J. B. (1997). High school size: Which works best and for whom? *Educational Evaluation and Policy Analysis*, *19*, 205-227. doi:10.3102 /01623737019003205

Leung, A., & Ferris, J. S. (2008). School size and youth violence. *Journal of Economic Behavior and Organization*, *65*, 318-333. doi:10.1016/j.jebo.2005.10.001

Lewis, W. C., &Chakraborty, K. (1996). Scale economics in public education. *Regional Analysis and Policy*, *26*, 23–35.

Lubienski, S. T., Lubienski, C., & Crane, C. C. (2008). Achievement differences and school type: The role of school climate, teacher certification, and instruction. *American*

Journal of Education, *115*, 97-138. doi:10.1086/590677

Luyten, H. (1994). School size effects on achievement in secondary education: Evidence from the Netherlands, Sweden, and the USA. *School Effectiveness and School Improvement*, *5*, 75-99. doi:10.1080/0924345940050105

Ma, X., & McIntyre, L. J. (2005).Exploring differential effects of mathematics courses on mathematics achievement. *Canadian Journal of Education/Revue Canadienne de l'éducation*, *28*, 827-852.

Maerten-Rivera, J., Myers, N., Lee, O., & Penfield, R. (2010). Student and school predictors of high-stakes assessment in science. *Science Education*, *94*, 937-962. doi:10.1002 /sce.20408

McMillen, B. J. (2004). School size, achievement, and achievement gaps. *Education Policy Analysis Archives*, *12*(58), 1-26. Retrieved from http://epaa.asu.edu/epaa/v12n58/

McNeal, R. B., Jr. (1999). Participating in high school extracurricular activities: Investigating school effects. *Social Science Quarterly, 80*, 291-309.

McNeely, C. A., Nonnemaker, J. M., & Blum, R. W. (2002).Promoting school connectedness: Evidence from the national *longitudinal* study of adolescent health. *Journal of School Health*, *72*, 138–146. doi:10.1111/j.1746-1561.2002.tb06533.x

Merkies, A. H. Q. M. (2000). Economics of scale and school consolidation in Dutch primary school industry. In J. L. T. Blank (Ed.), *Public provision and performance: Contributions from efficiency and productivity measurement* (pp. 191-218). Amsterdam, New York and Oxford: Elsevier Science, North-Holland.

Moe, T. M. (2009). Collective bargaining and the performance of the public schools. *American Journal of Political Science*, *53*, 156-174. doi:10.1111/j.1540-5907.2008 .00363.x

Mooij, T., Smeets, E., & De Wit, W. (2011). Multi-level aspects of social cohesion of secondary schools and pupils' feelings of safety. *British Journal of Educational Psychology*, *81*, 369-390. doi:10.1348/000709910X526614

O'Moore, A. M., Kirkham, C., & Smith, M. (1997). Bullying behavior in Irish schools: A nationwide study. *Irish Journal of Psychology*, *18*, 141-169. doi:10.1080/03033910.1997.10558137

Payne, A. A. (2012). Communal school organization effects on school disorder: Interactions with school structure. *Deviant Behavior*, *33*, 507-524. doi:10.1080/01639625.2011. 636686

Ready, D. D., & Lee, V. E. (2007). Optimal context size in elementary schools: Disentangling the effects of class size and school size and school size. *Brookings Papers on Education Policy* (pp. 99-135) Washington DC: Brookings Institution Press.

Rosenblatt, Z. (2001). Teachers' multiple roles and skill flexibility: Effects on work attitudes. *Educational Administration Quarterly*, *37*, 684-708. doi:10.1177/00131610121969479

Rumberger, R. W., & Palardy, G.J. (2005). Test scores, dropout rates, and transfer rates as alternative indicators of high school performance. *American Educational Research Journal*, *42*, 3-42. doi:10.3102/00028312042001003

Sandy, J., & Duncan, K. (2010). Examining the achievement test score gap between urban and suburban students. *Education Economics*, *18*, 297-315. doi:10.1080 /09645290903465713

Sawkins, J. W. (2002). Examination performance in Scottish secondary schools: An ordered logic approach. *Applied Economics*, *34*, 2031–2041. doi:10.1080/00036840210124559

Schneider, B. L., Wyse, A. E., & Keesler, V. (2006/2007). Is small really better? Testing some assumptions about high school size. *Brookings Papers on Education Policy* (pp. 15-47). Washington DC: Brookings Institution Press.

Silins, H., & Mulford, B. (2004). Schools as learning organisations - Effects on teacher leadership and student outcomes. *School Effectiveness and School Improvement*, *15*, 443-466. doi:10.1080/09243450512331383272

Stewart, E. A. (2003). School social bonds, school climate, and school misbehavior: A multilevel analysis. *Justice Quarterly*, *20*, 575-604. doi:10.1080/07418820300095621

Stewart, E. B. (2008). School structural characteristics, student effort, peer associations, and parental involvement The influence of school- and individual-level factors on academic achievement. *Education and Urban Society*, *40*, 179-204. doi:10.1177 /0013124507304167

Stiefel, L., Berne, R., Iatarola, P., & Fruchter, N. (2000). High school size: Effects on budgets and performance in New York City. *Educational Evaluation and Policy Analysis*, *22*, 27-39. doi:10.3102/01623737022001027

Stiefel, L., Schwartz, A. L., & Ellen, I. G. (2006).Disentangling the racial test score gap: probing the evidence in a large urban school district. *Journal of Policy Analysis and Management*, *26*, 7-30. doi:10.1002/pam.20225

Sun, L. T., Bradley, K. D., & Akers, K. (2012). A multilevel modelling approach to investigating factors impacting science achievement for secondary school students: PISA Hong Kong sample. *International Journal of Science Education*, *34*, 2107-2125. doi:10.1080 /09500693.2012.708063

Tanner, K. C., & West, D. (2011). *The effects of school size on academic outcomes*. Retrieved from http://sdpl.coe.uga.edu/research/SchoolSizeSDPL.pdf

Vegt, A. L van der., Blanken, M. den, & Hoogeveen, K. (2005). *Nationale scholierenmonitor: meting voorjaar 2005*. Utrecht: Sardes.

Vieno, A., Perkins, D. D., Smith, T. M., & Santinello, M. (2005). Democratic school climate and sense of community in school: A multilevel analysis. *American Journal of Community Psychology*, *36*, 327-341. doi:10.1007/s10464-005-8629-8

Watt, T. T. (2003). Are small schools and private schools better for adolescents' emotional adjustment? *Sociology of Education*, *76*, 344-367.

Wei, H. S., Williams, J. H., Chen, J. K., & Chang, H. Y. (2010). The effects of individual characteristics, teacher practice, and school organizational factors on students' bullying: A multilevel analysis of public middle schools. *Children and Youth Services Review*, *32*, 137-143. doi:10.1016/j.childyouth.2009.08.004

Weiss, C. C., Carolan, B. V., & Baker-Smith, E. C. (2010). Big school, small school: (Re)testing assumptions about high school size, school engagement and mathematics achievement. *Journal of Youth and Adolescence*, *39*, 163-176. doi:10.1007/s10964-009 -9402-3

Winter, M. de (2003).*Niet te groot en niet te klein: effecten van schaalgrootte op het welbevinden van jongeren*. Utrecht: NIZW.

Wyse, A. E., Keesler, V., & Schneider, B. (2008). Assessing the effects of small school size on mathematics achievement: A propensity score-matching approach. *Teachers College Record*, *110*(9), 1879-1900.

Table A1

Summary of the 46 studies (64 samples) of school size on student achievement used in the vote count

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Åberg-Bengtsson (2004)		Sweden	P	Rural schools: categories: < 75, > = 75			Lang	124		SEM	0	1	0	0	
Alspaugh (2004)		USA	P	Categories: < 200, 200-299, 300-399, 400-499, >= 500			GAA			A	0	0	1	0	<200 highest mean, 300-399 lowest
Archibald (2006)		USA	P	Number of students enrolled	548	137	Lang, Math			ML	2	0	0	0	
Barnes et al. (2006)	KS1	England	P	Total number of students at the school roll			Lang, Math			R	0	2	0	0	
	KS2						Lang, Math, Science				0	3	0	0	
Bickel et al. (2001)		USA	S	Number of students/1000	877	850	Lang, Math			R	0	3	0	0	
				In natural logarithms of single-student units			GAA				0	1	0	0	
Borland & Howsen (2003)		USA	P	Number of students	490	204	Lang & Math combined			R	0	0	0	1	
				School size squared			Lang & Math combined				0	0	1	0	∩ 760
Bowles & Bosworth (2002)		USA	PS	Average daily membership			Lang & Math combined	80		R	0	1	0	0	

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Bradley & Taylor (1998)	11-16	UK	S	Pupils/100			GAA			R					
	1992				685			1307			0	0	0	1	
	1996				765			1377			0	0	0	1	
	1992			Pupils/100 squared							0	0	1	0	∩ 1130
	1996										0	0	1	0	∩ 1230
	11-18			Pupils/100			GAA								
	1992				916			1580			0	0	0	1	
	1996				1010			1514			0	0	0	1	
	1992			Pupils/100 squared							0	0	1	0	∩ 1350
	1996										0	0	1	0	∩ 1440
Caldas (1993)	P	USA	P	Number of students enrolled	507	223	GAA	737		R	1	0	0	0	
	S		S		683	384	GAA	468			0	1	0	0	
Carolan (2012)		USA	S	Categories: < 600, 600-999, 1000-1599, >1599			Math	579		ML	0	1	0	0	
Chen & Weikart (2008)		USA	S	Number of students enrolled	960	493	Lang & Math combined	212		SEM	0	1	0	0	
Coladarci & Cobb (1996)		USA	S	Categories: <800, >=1600			Lang & Math combined		4567	SEM	0	1	0	0	
Deller & Rudnicki (1993)		USA	P	Average daily attendance			GAA	139		R	1	0	0	0	
Driscoll et al. (2003)	Primary	USA	P	School size	526	394	GAA	4025		R	1	0	0	0	
	Middle		S		526	394		753			0	1	0	0	
	High school		S		526	394		747			0	1	0	0	

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Duran-Narucki (2008)		USA	P	Number of students enrolled	712	328	Lang & Math combined (poor achievers)	95		R	0	2	0	0	Attendance is mediator
Eberts et al. (1990)		USA	P	Categories: < 200, 400-599, >800			Math		1400	R	2	0	0	0	
Fernandez (2011)		USA	PS	The number of students enrolled	1082	637	Lang & Math combined	252		R	0	2	0	0	
Foreman-Peck & Foreman-Peck (2006)	1996 2002	UK	S	Ln (previous year pupil numbers)	871 936	331 519	GAA	1119		LR	0	0	0	1	
				Ln school size squared							0	0	1	0	∩ 560
Fowler & Walberg (1991)		USA	S	Total enrolment	1070	519	Lang & Math	293		R	5	8	0	0	
Gardner et al. (2000)		USA	S	Categories: 200-600 vs > 2000	424 2500		Lang & Math			A	0	1	0	1	
Heck (1993)		USA	PS	Actual size of enrolment			Lang & Math	235		R	2	0	0	0	
Holas & Huston (2012)	Grade 5	USA	PS	Total enrolment	490	210	Lang & Math	10	855	SEM	0	2	0	0	
	Grade 6				690	300					0	1	0	0	
Kahne et al. (2008)		USA	S	School size			Lang & Math	80		ML	0	4	0	0	
Kuziemko (2006)		USA	P	Abrupt change in school enrolment	418	170	Math	>100		R	0	6	0	0	

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Lamdin (1995)		USA	P	Number of students enrolled	469	172	Lang & Math	97		R	0	6	0	0	
Lee et al. (2011)	03-04 04-05 05-06 06-07 07-08	USA	S	>800 vs small learning communities (400 students)			GAA	>230		T	1	4	0	0	
Lee & Loeb (2000)		USA	P	Categories: <400, 400-750, >750			Math	264	4495	ML	1	1	0	0	
Lee & Smith (1995)		USA	S	Ln total enrolment			Lang, Math, Science, Other	820	11794	ML	4	0	0	0	
Lee & Smith (1997)		USA	S	Categories: <300, 301-600, 601-900, 901-1200, 1201-1500, 1501-1800, 1801-2100, >2100			Lang & Math	789	9812	ML	0	0	2	0	∩601-900 ∩601-900
Lubienski et al. (2008)	Grade 4	USA	P	Categories: 1-299, 300-499, 500-699,>=700			Math		157161	ML	0	1	0	0	
	Grade 8		S	Categories: 1-399, 400-599, 600-799, 800-999, >=1000					119364		0	0	0	1	
Luyten (1994)	USA 1st and 2nd sample	USA	S	Categories: <240, 240-359, 360-499, 500-999, >1000			Math	58	2212	ML	0	1	0	0	

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
	Sweden	Sweden					Math	95	3500		0	1	0	0	
	Nether-lands (Math)	Nether-lands (Math)					Math	228	5313		0	1	0	0	
	Nether-lands (Science)	Nether-lands (Science)					(Earth) Science	194	4286		0	1	0	0	
Ma & McIntyre (2005)		Canada	S	Expressed in hundred student units	613	363	Math	34	1518	ML	0	1	0	0	
Maerten-Rivera et al. (2010)		USA	P	School size	798	331	Science	198	23854	ML	0	1	0	0	
McMillan (2004)	Primary school	USA	P	Categories: <400, 400-549, 550-699, >700	506		Lang & Math	1053		ML	0	2	0	0	
	Middle school		S	Categories: <400, 400-549, 550-699, >700	570		Lang & Math	508			0	2	0	0	
	High school		S	Categories: <700, 700-1199, 1200-1699, >1700	859		Lang & Math	333			0	0	0	2	
Moe (2009)	Primary	USA	P	The log of school enrolment			GAA	1947		R	1	0	0	0	
	Secon-dary		S	The log of enrolment			GAA	829			0	1	0	0	
Ready & Lee (2006)		USA	P	Categories: <275, 276-400, 601-800 (RF), 601-800, >800			Lang & Math	527	7740	ML	0	2	0	0	

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Rumberger & Palardy (2005)		USA	S	Categories: 1-600, 601-1200 (RF), 1201-1800, >1800			GAA	912	14199	ML	0	0	1	0	∩1200-1800
Sandy & Duncan (2010)	Urban	USA	S	<1000 vs >1000			GAA		1955	R	0	1	0	0	
	Sub-urban										0	1	0	0	
Sawkins (2002)	1993-1994	UK (Scotland)	S	Total number of pupils/100	796	356	GAA	398		R	1	0	0	0	
				(Total number of pupils/100) squared							0	0	1	0	U1190
	1998-1999			Total number of pupils/100	806	356					1	0	0	0	
				(Total number of pupils/100) squared							0	0	1	0	U1230
Schneider et al. (2006/2007)		USA	S	Categories: 1-399, 400-799, 800-119 (RF), 1200-1999, >=2000			Math	660	12489	ML	0	1	0	0	
Stewart (2008)		USA	S	Total student enrolment	1540	686	GAA	715	11999	ML	0	1	0	0	
Stiefel et al. (2006)	Grade 5	USA	P	Enrolment Subgroups: Asian, Black, Hispanic, White	958		Lang	667	70638	ML	1	0	0	0	
	Grade 8		S	Subgroups: Asian, Black, Hispanic, White	1221			278	55921		0	1	0	0	

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Sun et al. (2012)		Hong Kong	S	Total school enrolment	1039	174	Science	145	4645	ML	0	0	0	1	
Tanner & West (2011)		USA	S	Net enrolment	1370	682	Lang, Math, Science, Other	303		R	0	6	0	0	
Weiss et al. (2010)		USA	S	Categories: 1-599 (RF), 600-999, 1000-1599, 1600-2499			Math		10946	ML	0	1	0	0	
Wyse at el. (2008)		USA	S	Categories: 1-399, 400-799, 800-1199, 1200-1999, >=2000			Math	745	12853		0	1	0	0	

P: primary education, S: secondary education, Ln: Natural Logarithm, Lang: language, GAA = general academic achievement (composite), A: An(c)ova, LR: Logistic Regression, P: Pearson correlation analysis, R: Regression analysis, ML: Multilevel analysis, SEM: Structural Equation Modeling, T=T-test

- = negatively related with school size, ns = no significant relation with school size, ∩ = optimal school size found, += positively related with school size

Table A2

Summary of the 14 studies (15 samples) of school size on students' and teachers' attitudes towards schools) used in the vote count

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Bowen et al. (2000)		USA	S	Categories: 0-399, 400-599, 600-799, 800-999, 1000-1399	689		School satisfaction	39	945	A	1	0	0	0	
							Teacher support								
Crosnoe et al. (2004)		USA	S	Enrolment/100	1381	838	School attachment	84	13162	ML	1	0	0	0	
							Student-teacher bonding								
				(Enrolment/100)^2							0	0	1	0	U1900-2000
Holas & Huston (2012)	Grade 5	USA	P	Total enrolment	490	210	School attachment		827	SEM	0	1	0	0	
Kahne et al. (2008)		USA	S	School size			Teacher-teacher trust	80		ML	0	1	0	0	
							Academic personalism				1	0	0	0	
							Classroom personalism				1	0	0	0	
							Sense of belonging				1	0	0	0	
							Student-teacher trust				1	0	0	0	
							Teacher support				1	0	0	0	

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Kirkpatrick Johnson et al. (2001)	Middle schools	USA	S	Total enrolment/100	477	234	School attachment	45	2482	ML	0	1	0	0	
	High Schools				1147	716		64	8104		0	1	0	0	
Koth et al. (2008)		USA	P	School enrolment			Achievement motivation	37	2468	ML	1	0	0	0	
Lee & Loeb (2000)		USA	P	Categories: <400 (RF), 400-750, >750			Teachers' collective responsibility	264	22599	ML	2	0	0	0	
McNeely (2002)		USA	S	Ln school size (in 100s)	642	765	School connectedness	127	75515	ML	1	0	0	0	
Payne (2012)		USA	S	Ln student enrolment	792	479	Communal school organization	253		R	0	1	0	0	
Rosenblatt (2001)		Israel	S	Number of students/100	1020	650	Organizational commitment	12		SEM	1	0	0	0	
Silins & Mulford (2004)		Australia	S	Size	632	283	Students' engagement with school	96	3500	SEM	1	0	0	0	
Vegt et al. (2005)		Netherlands	S	Number of pupils at school site			School connectedness	51		R	0	1	0	0	
							Relationships with peers				1	0	0	0	
							Relationship with teachers				1	0	0	0	

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Vieno et al. (2005)		Italy	S	Size of student body	480	304	Students' sense of community	134	248	ML	0	1	0	0	
Winter, de (2003)		Nether-lands	S	Categories: <500, 500-1000, >1000			Classroom climate			A	0	0	1	0	∩500-1000

P: primary education, S: secondary education, Ln: Natural Logarithm, A: An(c)ova, P: Pearson correlation analysis, R: Regression analysis, ML: Multilevel analysis, SEM: Structural Equation Modeling, T=T-test

- = negatively related with school size, ns = no significant relation with school size, ∩ = optimal school size found, += positively related with school size

Table A3

Summary of the 10 studies (10 samples) of school size on participation of students, teachers or parents used in the vote count

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Coladarci & Cobb (1996)		USA	S	Categories: <800) vs >=1600			Extra-curricular participation		4567	SEM	1	0	0	0	
Crosnoe et al. (2004)		USA	S	Enrolment/100	1381	838	Student extra-curricular participation	84	13420	ML	1	0	0	0	
Dee at el. (2007)		USA	S	Categories: <400 (RF), 400-799, 800-1199, 1200-2199, >=2200			Parents act as volunteer at school	390	8197	ML	0	1	0	0	
Feldman & Matjasko (2006)		USA	S	Categories: 1-400 (RF), 401-1000, >1000			Adolescent extra-curricular participation	132	13810	LR	1	0	0	0	
Gardner et al. (2000)		USA	S	Categories: 200-600 vs >2000 pupils	424 2500		Average parent teacher association members for each school	127		A	1	0	0	0	
Holas & Huston (2012)	Grade 6	USA	PS	Total enrolment	690	300	School involvement		825	SEM	1	0	0	0	
Kahne et al. (2008)		USA	S	School size			Teacher influence	80		ML	1	0	0	0	
Lay (2007)		USA	S	Continuous measure			Participation in school activities		3010	LR	1	0	0	0	

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
				Categories: <300, 301-600, 601-900, 901-1200, 1201-1500, 1501-1800, >1800							0	0	1	0	U1500-1800
				Categories: <300, 300-599, 600-999, >1000							0	1	0	0	
McNeal (1999)		USA	S	Ln number of students	1053		Student participation in school activities	281	5772	ML	1	0	0	0	
							Athletics			LR	1	0	0	0	
Silins & Mulford (2004)		Australia	S	School size	632	283	Student (extra-) curricular participation	96	3500	SEM	1	0	0	0	

P: primary education, S: secondary education, Ln = Natural logarithm, A: An(c)ova, LR = Logistic Regression, P: Pearson correlation analysis, R: Regression analysis, ML: Multilevel analysis, SEM: Structural Equation Modeling, T=T-test

- = negatively related with school size, ns = no significant relation with school size, ∩ = optimal school size found, + = positively related with school size

Table A4

Summary of the 24 studies (25 samples) of school size on safety used in the vote count

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Attar-Schwartz (2009)		Israel	S	Number of students	557	332	Sexual harassment victimization	327	16604	ML	1	0	0	0	
Bonnet et al. (2009)		Nether-lands	P	Categories: <300, 301-500, >500			Victimization	23	2003	ML	0	1	0	0	
Bowen et al. (2000)		USA	S	Categories: 0-399, 400-599, 600-799, 800-999, 1000-1399	689		School safety	39	945	A	1	0	0	0	
Bowes et al. (2009)		England	P	Total number of children	291	136	Involvement in bullying		2232	LR	1	2	0	0	
Chen (2008)		USA	S	Categories: <300 , 300-499, 500-999, >=1000			Misbehavior			SEM	1	0	0	0	
							Crime incidents				1	0	0	0	
Chen & Weikart (2008)		USA	S	Number of students enrolled	960	493	School disorder	213		SEM	0	1	0	0	
Chen & Vazsony (2012)		USA	S	Categories: <400, 400-1000, >1000			Problem behavior	85	9163	ML	1	0	0	0	
Eccles et al. (1991)		USA	P&S	Total school enrolment			Violence	759		R	1	0	0	0	
							Substance abuse while at school				1	0	0	0	

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Gottfredson & DiPietro (2011)		USA	S	Ln number of students	792	478	Property victimization	253	13597	ML	0	1	0	0	
							Personal victimization				1	0	0	0	
Haller (1992)		USA	PS	Enrolment	963	1219	Disorder reported by: principals	558		R	1	0	0	0	
							Students				0	1	0	0	
							Self				1	0	0	0	
Heck et al. (1993)		USA	PS	Enrolment			Substance abuse	235		R	0	1	0	0	
Inspectorate of Education (2003)		Nether-lands	S	Categories: <500, 501-1000, >1000			Pupil guidance and school climate	378		A	0	1	0	0	
Kahne et al. (2008)		USA	S	School size			Respectful classroom behavior	80		ML	0	1	0	0	
Khoury-KassabrI et al (2004)		Israel	S	Number of students	505	298	Victimization: physical	162	10400	ML	0	1	0	0	
							Threats				0	1	0	0	
							Moderate physical violence				0	1	0	0	
							Verbal-social				0	1	0	0	
Klein & Cornel (2010)		USA (Virginia)	S	School enrolment size	1210	690	Self-report bully victimization	290	7431	R	0	1	0	0	

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
							Student perceptions of bullying				1	0	0	0	
							Teacher perceptions of bullying				1	0	0	0	
							Total bullying violations				1	0	0	0	
							Bullying violations rate				0	0	0	1	
							Self-report threat victimization				1	0	0	0	
							Total threat violations				1	0	0	0	
							Threat violations rate				0	0	0	1	
							Self-report physical victimization				0	1	0	0	
							Total attack violations				1	0	0	0	
							Attack violations rate				0	0	0	1	
Koth et al. (2008)		USA	P	School enrolment			Order and discipline	37	2468	ML	0	1	0	0	
Leung & Ferris (2008)		Canada	S	Number of students/1000			Youth violence	110	616	LR	1	0	0	0	

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
				Categories: <999 (RF), 1000-1499, 1500-1999, >2000							0	1	0	0	
Mooij et al. (2011)		Nether-lands	S	Number of pupils/100	926	514	Pupils' feelings of safety at school	104	26162	ML	0	0	0	1	
O'Moore et al. (1997)	P	Ireland	P	Categories: 0-199, 200-499, >=500			Being bullied	320	9559	A	0	1	0	0	
							Bullying				0	0	1	0	∩200-499
	S		S	Categories: 0-199, 200-499, >=500			Being bullied				0	0	1	0	U>=500
							Bullying				0	0	0	1	
Stewart (2003)		USA	S	School enrolment	1540	686	School misbehavior	528	10578	ML	1	0	0	0	
Vegt et al. (2005)		Nether-lands	S	Number of pupils at school site			Safety	51	5300	R	0	1	0	0	
							Safety policy				0	0	0	1	
							Bullying and fighting				1	0	0	0	
							Vandalism				1	0	0	0	
Watt (2003)	Males Females	USA	S	Categories: <=400, 401-1000, 1001-4000			Violence		12150	LR	0	1	0	0	
Wei et al. (2010)		Taiwan	S	Total number of students	1586	989	Physical bullying	12	1172	ML	0	1	0	0	

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
							Verbal bullying				0	1	0	0	
Winter (2003)		Nether-lands	S	Categories: <500, 500-1000, >1000			Being bullied		5726	A	0	0	0	1	
							Bullying				0	0	0	1	
							Fighting				0	0	0	1	

P: primary education, S: secondary education, Ln: Natural Logarithm, A: An(c)ova, LR: Logistic Regression, P: Pearson correlation analysis, R: Regression analysis, ML: Multilevel analysis, SEM: Structural Equation Modeling, T=T-test

- = negatively related with school size, ns = no significant relation with school size, ∩ = optimal school size found

Table A5

Summary of the 12 studies (19 samples) of school size on attendance and truancy used in the vote count

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Bos et al. (1990)		Nether-lands	S	School size			Truancy			R	0	1	0	0	
Chen & Weikart (2008)		USA	S	Number of students enrolled	960	493	Attendance rate	213		SEM	1	0	0	0	
Duran-Narucki (2008)		USA	P	The number of students enrolled	712	328	Attendance	95		R	0	0	0	1	
Eccles et al. (1991)		USA	P&S	Total school enrolment			Absenteeism	759		R	1	0	0	0	
Foreman-Peck & Foreman-Peck (2006)		UK	S	Log (previous year pupil numbers) 1996	871	331	% of non-attendance	1119		LR	1	0	0	0	
				2002	936	328									
Gardner et al. (2000)		USA	S	Categories: 200-600 vs >2000	424 2500		Absenteeism rate	127		A	1	0	0	0	
Haller (1992)		USA	PS	Enrolment	963	1219	Truancy reported by principals	558		R	1	0	0	0	
							Students				1	0	0	0	
							Self				1	0	0	0	
Heck (1993)		USA	PS	Actual size of enrolment			Attendance	235		R	1	0	0	0	
Jones et al. (2008)		USA	S	School enrolment	1012	849	Attendance	1039		R	1	0	0	0	

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Kahne et al. (2008)	2002-2003	USA	S	School size			Absences	80		ML	1	0	0	0	
	2003-2004										0	1	0	0	
	2004-2005										0	1	0	0	
	2005-2006										1	0	0	0	
Kuziemko (2006)		USA	P	Abrupt change in school enrolment	418	170	Change in average daily attendance			R	0	2	0	1	
Lee et al. (2011)	2003-2004	USA (Ohio)	S	Small schools (<400) vs. traditional schools (>800)			Attendance rate			M-W	0	1	0	0	
	2004-2005										0	1	0	0	
	2005-2006										0	1	0	0	
	2006-2007										0	1	0	0	
	2007-2008										1	0	0	0	

P: primary education; S: secondary education; A: An(c)ova, LR: Logistic Regression, P: Pearson correlation analysis, R: Regression analysis; ML: Multilevel analysis, M-W: Mann-Whitney test, SEM: Structural Equation Modeling, T=T-test

- = negatively related with school size, ns = no significant relation with school size, ∩ = optimal school size found, + = positively related with school size

Table A6

Summary of the 4 studies (5 samples) of school size on drop-out used in the vote count

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Gardner et al. (2000)		USA	S	Categories: 200-600 vs. >2000	424 2500		Dropout rate	127		A	1	0	0	0	
Kahne et al. (2008)	2002-2003	USA	S	School size			Dropout rate	80		ML	0	1	0	0	
	2003-2004										0	1	0	0	
Lee & Burkam (2003)		USA	S	Categories: 0-600, 601-1500 (RF), 1501-2500			Dropout rate	190		ML	0	0	1	0	U601-1500
Rumberger & Palardy (2005)		USA	S	Categories: 1-600, 601-1200 (RF), 1201-1800, >1800			Dropout rate	912	14199	ML	0	0	1	0	∩1200-1800

P: primary education, S: secondary education, Ln: Natural Logarithm, A: An(c)ova, P: Pearson correlation analysis, R: Regression analysis, ML: Multilevel analysis, SEM: Structural Equation Modeling, T=T-test

- = negatively related with school size, ns = no significant relation with school size, ∩ = optimal school size found, + = positively related with school size

Table A7

Summary of the 5 studies (6 samples) of school size on other student outcome variables (attitudes towards self and learning, engagement) used in the vote count

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Coladarci & Cobb (1996)			S	Categories: <800 vs >=1600			Self-esteem		4567	SEM	0	1	0	0	
Holas & Huston (2012)	Grade 6	USA	P	Total enrolment	690	300	Self-competence		828	SEM	0	1	0	0	
Kirkpatrick Johnson et al. (2001)	Middle schools	USA	S	Total enrolment/100	477	234	Engagement in school	45	2482	ML	0	1	0	0	
	High schools				1147	716		64	8104		1	0	0	0	
Lay (2007)		USA	S	Continuous measure			Participation in community services		3010	LR	0	1	0	0	
				Categories: <300, 301-600, 601-900, 901-1200, 1201-1500, 150 1800, >1800							0	0	1	0	∩<300
				Categories: <300, 300-599, 600-999, >1000							0	0	1	0	∩<300
Lee & Smith (1995)		USA	S	Ln total enrolment			Academic engagement	820	11794	ML	1	0	0	0	
Weiss et al. (2010)		USA	S	Categories: 1-599 (RF), 600-999, 599, 1600-2499			School engagement		10946	ML	1	0	0	0	

P: primary education, S: secondary education, Ln: Natural Logarithm, A: An(c)ova, LR: Logistic Regression, P: Pearson correlation analysis, R: Regression analysis, ML: Multilevel analysis, SEM: Structural Equation Modeling, T=T-test

- = negatively related with school size, ns = no significant relation with school size, ∩ = optimal school size found, + = positively related with school size

Table A8

Summary of the 4 studies (4 samples) of school size on school organization and teaching and learning used in the vote count

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Eccles et. al (1991)		USA	P&S	Total school enrolment			Teacher efficacy	759		R	1	0	0	0	
Inspectorate of Education (2003)		Nether-lands	S	Categories: <500, 501-1000, >1000			Pedagogic and didactic approac	378		A	0	0	1	0	U500-1000
Kahne et al. (2008)		USA (Chicago)	S	School size			Collective responsibility	80		ML	1	0	0	0	
							Commitment to innovation				1	0	0	0	
							Expectations post-secondary education				1	0	0	0	
							Principal instructional leadership				0	1	0	0	
							Program coherence				0	1	0	0	
							Quality professional development				0	1	0	0	
							Quality student discussions in classroom				0	1	0	0	
							Reflective dialogue				0	1	0	0	
							Academic press				0	1	0	0	
							Quality English				0	1	0	0	

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
							instruction								
							Quality Math instruction				0	1	0	0	
							Peer support for academic achievement				1	0	0	0	
							School-wide future orientation				0	1	0	0	
Silins & Mulford (2004)		Australia	S	School size in 1997	632	283	Organisational learning	96		SEM	0	1	0	0	
							Teacher leadership				1	0	0	0	
							Teachers’ work the classroom				0	1	0	0	

P: primary education, S: secondary education, A: An(c)ova, Ln: Natural Logarithm, P: Pearson correlation analysis, R: Regression analysis, ML: Multilevel analysis, SEM: Structural Equation Modeling, T=T-test

- = negatively related with school size, ns = no significant relation with school size, ∩ = optimal school size found, + = positively related with school size

Table A9

Summary of the 5 studies (5 samples) of school size on costs used in the vote count

Authors (publication year)	Sample	Countries in sample	School type	Measure of school size	Mean	SD	Outcome measure	Schools (N)	Students (N)	Statistical technique employed	Direction of the effect				Comments
											-	ns	∩	+	
Bickel et al. (2001)		USA	S	Ln number of students	877	850	Expenditure per pupil	1001		R	1	0	0	0	
Bowles & Bosworth (2002)		USA	PS	Ln average daily membership			Ln Expenditure per pupil	80		R	1	0	0	0	
Lewis & Chakraborty (1996)		USA	PS	Ln number of students	511		Ln Expenditure per pupil			R	1	0	0	0	Mean is median
Merkies (2000)		Nether-lands	P	Ln number of pupils	200		Ln Expenditure per pupil	1784		R	1	0	0	0	
Stiefel et al. (2000)		USA	P	Ln school size	2030	1192	Ln Expenditure per puil	121		R	0	1	0	0	

P: primary education, S: secondary education, Ln: Natural logarithm, A: An(c)ova, P: Pearson correlation analysis, R: Regression analysis, ML: Multilevel analysis, SEM: Structural Equation Modeling, T=T-test

- = negatively related with school size, ns = no significant relation with school size, ∩ = optimal school size found, + = positively related with school size

3

School leadership effects revisited; A review of empirical studies guided by indirect-effect models[1]

[1] This chapter is based on:
Hendriks, M.A. & Scheerens, J. (2013). School leadership effects revisited: a review of empirical studies guided by indirect-effect models, *School Leadership & Management*, 33, 373-394. doi:10.1080/13632434.2013.813458

Abstract

Fifteen leadership effect studies that used indirect-effect models were quantitatively analyzed to explore the most promising mediating variables. The results indicate that total effect sizes based on indirect-effect studies appear to be low, quite comparable to the results of some meta-analyses of direct-effect studies. As the earlier indirect-effect studies tended to include a broad range of mainly school organizational conditions as intermediary variables, more recent studies focus more sharply on instructional conditions. The results of the conceptual analysis and the quantitative research synthesis would seem to support conceptualizing educational leadership as a detached and 'lean' kind of meta-control, which would make maximum use of the available substitutes and self-organization offered by the school staff and school organizational structural provisions. The coupling of conceptual analysis and systematic review of studies driven by indirect-effect models provides a new perspective on leadership effectiveness.

Conceptualization

As an introduction to the review and quantitative synthesis of school leadership effects the development of leadership concepts and research models will be briefly reviewed.

From Direct to Indirect-Effect Models

Earlier reviews of leadership effects were based on so-called *'direct' effects* of leadership on student performance outcomes (e.g. Witziers, Bosker & Krüger, 2003). Basically, simple correlations between leadership characteristics and student achievement, sometimes adjusted for student background characteristics, were at the focus of these reviews. Despite the presence of conceptual 'indirect' models of leadership effects, in which the influence of leadership is seen as 'meditated' by school-level conditions, like the model by Bossert, Dwyer, Rowan and Lee (1982), empirical studies that were guided by these models took some time to be realized: one of the first well-known empirical studies of this nature being the work of Mulford (2003) and Silins and Mulford (2004), the LOSLO project. Characteristics of the school organization, the school climate and perceptions of teachers' work are examples of the intermediary school conditions addressed in these models. Substantively, it makes a lot of sense to see the influence of school leadership behavior on student outcomes as indirect, and as the result of a hypothetical causal chain, through which school heads directly operate on school organizational and instructional conditions, which in their turn influence student achievement. Methodological tools to study such indirect-effect models are available in path analysis and structural equation modelling. Earlier reviews (e.g. Witziers, Bosker & Krüger, 2003) could only refer to a very limited number of 'indirect effect' studies. In this article we are presenting a quantitative overview of effect sizes of a somewhat larger (but still limited) number of recent studies based on indirect-effect models.

From Instructional Leadership to Integrated Leadership

The conceptual development of school leadership can be interpreted as a gradual broadening of the construct of educational leadership, starting out from a very focused concept of instructional leadership, in the early school effectiveness studies by, among others Edmonds (1979), to a more encompassing perspective, indicated as integrated leadership.

The picture of the instructional leader that emerged from early school effectiveness research is that of the leader as a facilitator and controller of the primary process of teaching and learning. The appearance of such concepts as 'curricular leadership' (Glatthorn, 1987, 1997) and 'instructional leadership' (Hallinger, 1983) reflects this image. This focused view of instructional leadership is concentrated on only one of the four core leadership practices mentioned by Leithwood, Jantzi and McElheron-Hopkins (2006), namely managing the instructional program. In a next phase, these relatively narrow conceptualizations of instructional leadership were broadened gradually; actions and strategies other than those closely related to the primary process of teaching and learning come into view as well. Hallinger (1983), for example, stated that instructional leadership is related to *defining a mission for the school, managing curriculum and instruction* and *promoting a learning climate favorable for student learning*. One could see this perspective as an extended version of instructional leadership.

The concept of transformational leadership underlines the role of the school leader in promoting school improvement. This implies a focus on what one could call 'secondary processes' in shaping organizational structure and culture and capacity building of the staff. This concept emphasizes that one of the main tasks of school leaders is to initiate processes and structures, within the school, that enable teacher collaboration and participative decision-making. The concept is fuelled by the notion that, in many schools, teachers are autonomous and isolated, implying that school leaders should not intervene directly with curricular and instructional affairs, but rather indirectly by transforming the school culture to facilitate collegial planning, collaboration and experimentation aimed at school improvement. In other words, the main tasks of the school leader should be to create a working environment in which teachers collaborate and identify themselves strongly with the school's mission.

One facet of transformational leadership is empowerment of teachers and participative decision-making. This element comes close to the idea of distributed leadership, where teachers as autonomous professionals are seen as carrying out leadership tasks.

Finally the term 'integrated leadership' has been coined by several authors, e.g. Leithwood (1992).

Here we shall take a liberal interpretation of integrated leadership and take it as:

- an integration of transformational and instructional leadership, following Marks and Printy's (2003) argumentation that in order to be effective, transformational leadership aimed at school reform requires an additional component directed at teaching and learning;

- more encompassing leadership models, such as developed by authors like Bolman and Deal (1997) and Quinn, Faerman, Thompson & McGrath (1996), that include a broad perspective of organizational effectiveness;
- a view of leadership that emphasizes the distribution of coordination and leadership tasks not just among people (as in distributed leadership), but also to structural characteristics of the school organization. This perspective has been introduced by Heck and Hallinger (2009), as they discuss dynamic theories of organizational processes: 'Dynamic theories of organizational processes seek to describe how changes in organizational structures (e.g., size, hierarchy, staffing) and social-cultural interactions (e.g., organizational culture, decision-making structures, leadership, social networks) influence organizational outcomes over a period time)' (Hallinger & Heck, 2009, p. 105).

A schematic overview of the development in the concept formation on school leadership is presented in Table 3.1.

Table 3.1
Concepts of leadership at school

Type of leadership	**Orientation**
Instructional leadership	Curriculum and instruction
Extended instructional leadership	School mission Managing the curriculum Providing learning climate
Transformational leadership	Models organizational values Develops shared mission Provides intellectual stimulation Builds consensus Redesigns organizational structure
Integrated leadership	Conditions supporting school improvement Instructional leadership; broader perspectives on organizational effectiveness; leadership roles "delegated" to people and structural coordination mechanisms

Relationship between Indirect-Effect Models and School Leadership Concepts

The orientation of the different leadership types points at different school conditions that could be seen as intermediary variables in indirect leadership effect models. As the right hand column of Table 3.1 illustrates, the definition of the various leadership concepts is based on reference to functional domains of the school as an organization, such as the curriculum, organizational structures, the school climate, the capacity of teachers and so on.

Further elaboration is facilitated by the way Leithwood et al. (2006) summarize core functions of school leadership:

1. *Developing a vision and giving direction*: identifying and formulating a vision, creating a shared interest, high expectations of performance, promoting the acceptance of group objectives, monitoring organizational performance and communicating.
2. *Understanding and developing people*: providing intellectual stimulation, giving individual guidance and setting an example. The school leader builds on the knowledge and skills of teachers and other personnel to achieve the school objectives.
3. *Redesigning the organization:* building on cultures and cooperative processes, managing the environment and working conditions, building and maintaining productive relations with parents and the community, connecting the school with the wider environment.
4. *Managing the teaching and learning program*: creating a productive working environment for both teachers and students, promoting organizational stability, guaranteeing effective leadership with the focus on learning, appointing teachers and supporting staff to implement the curriculum, monitoring school activities and performance.

In recent empirical studies guided by indirect-effect models, a wide range of intermediary variables is being addressed. Theoretically one might expect that, for the choice of intermediary variables, the knowledge base of educational effectiveness would be used (Creemers & Kyriakides, 2008; Scheerens, 2000; Scheerens, Luyten, Steen & Luyten-de Thouars, 2007). In school and instructional effectiveness studies, variables are identified that have an effect on student achievement. As most of these variables can be influenced by school leaders, they are good candidates for being included in indirect-effect studies of school leadership.

Table 3.2 matches core leadership functions and educational effectiveness enhancing conditions at school and classroom level.

The table shows that there is a fair match between emphases in school leadership and school factors that have been empirically supported for being positively associated with student achievement. Accordingly, Table 3.2 provides a conceptual map for conducting leadership effect studies guided by indirect models. Examination of actual empirical research studies will reveal to what extent this framework has also been followed.

Table 3.2
Leadership functions, leadership behaviors and effectiveness enhancing school conditions

Leadership functions	**Leadership behavior**	**Effectiveness enhancing factors**
Developing a vision	External contacts Buffering Setting values	Enhanced teaching time Shared sense of purpose among teachers High expectations
Managing the teaching and learning program	Direction setting (vision, goals, standards Monitors curriculum and instruction (managing the instructional program) Redesigning the organization	Clear goals and standards Opportunity to learn Student monitoring & feedback procedures Structured teaching Active teaching Active learning
Understanding and developing people	HRM & HRD Coaches teachers Recruits teachers Builds consensus Individual support Intellectual stimulation	Cohesion among teachers Professionalization Teacher competency Teachers' sense of self efficacy
Redesigning the organization	Uses 'substitutes' for leadership Distributes leadership tasks Creates climate	Student monitoring & feedback provisions Disciplinary climate Supportive climate

The causal assumption in Table 3.2 is that the intentions and behavioral directions of school leaders, targeted at specific domains of school functioning, influence the way these domains are actually functioning and, in their turn influence student achievement. The empirical test of these theoretical assumptions obviously requires a research design that allows for causal interpretation, including the requirement that leadership behavior and effectiveness enhancing intermediary school conditions are measured independently.

Results of Earlier Meta-Analyses

Results

Meta-analyses provide a quantitative synthesis of research results from individual studies. In meta-analysis the findings are reported in terms of effect sizes (representing both the direction and magnitude of the effect). Although there are many different types of effect size measures, two main types often used are correlations and standardized mean differences. In school effectiveness research, effect sizes reflect the association of a particular effectiveness enhancing variable (in our case leadership) with an effect measure, like the results on a cognitive achievement test. In the context of effectiveness research these associations are usually rendered as correlations (indicated with the coefficient *r,* expressing the product moment correlation). In the research literature at large, effect sizes are often expressed as the standardized mean difference between an experimental and a control group (indicated with coefficient Cohen's *d*). The two coefficients (*r* and *d*) are convertible to one another[2]. In this article we refer to effect sizes using both of these coefficients, although we have consistently used effect sizes expressed in correlations in the tables.

In Table 3.3, the results of nine meta-analyses are summarized. These meta-analyses are by Scheerens and Bosker (1997), Witziers, Bosker and Krüger (2003), Marzano, Waters and McNulty (2005), Chin (2007), two by Robinson, Lloyd and Rowe (2008), Creemers and Kyriakides (2008), Scheerens et al. (2007), and Hattie (2009).[3]

Table 3.3
Summary of results from meta-analyses on school leadership; effect sizes are rendered as correlations between school leadership and student achievement

Meta-analysis by:	Leadership concept	Effect size (correlation)
Scheerens & Bosker (1997)	School leadership	*r* = .04
Witziers, Bosker & Krüger (2003)	School leadership	*r* = .02
Marzano, Waters & McNulty (2005)	Generalized school leadership	*r* = .25
Chin (2007)	Transformational leadership	*r* = .49
Robinson, Lloyd & Rowe (2008) (1)	Instructional leadership	*r* = .21
Robinson, Lloyd & Rowe (2008) (2)	Transformational leadership	*r* = .06
Creemers & Kyriakides (2008)	School leadership	*r* = .07
Scheerens et al. (2007)	School leadership	*r* = .06
Hattie (2009)	School leadership	*r* = .18

The average effect size across these meta-analyses comes down to *r* = 0.15. When leaving out the outlying value of the meta-analysis by Chin the average effect size would become *r* = 0.11.

In our own work we consistently find effect sizes in the order of *r* = 0.05, while some other meta-analyses have found much higher effect sizes.

[2] Converting from *r* to *d* (Borenstein, Hedges, Higgins & Rothstein, 2009, p. 48), is as follows:

$$\mathrm{d} = \frac{2\mathrm{r}}{\sqrt{1-\mathrm{r}^2}}$$

[3] Hattie presents effect sizes in terms of the standardized mean difference between experimental and control group, which are roughly twice the size of the correlation coefficient for low to medium effect sizes.

Interpretation of Effect Sizes

According to Cohen's standards for interpreting effect sizes[4], our results on leadership effects should be interpreted as negligible to small. It should be noted however, that several authors argue that Cohen's standards are to be considered as too conservative, and do not match the practical significance of malleable school variables. Richard, Bond and Stokes-Zoota (2003; cited by Baumert, Lüdtke and Trautwein, 2006) found a mean correlation of r = 0.21 in their meta-analysis of meta-analyses in social psychology, and proposed a modification of Cohen's classification, considering a correlation of 0.30 to indicate a large effect (p. 339). Baumert et al. (2006) propose the learning gain during one school year as a realistic standard to express effects of schooling. They cite several studies indicating that this learning gain has the magnitude of about d = 0.30, which would be comparable to a correlation of 0.15. These authors also discuss a method to compute effect sizes developed by Tymms, Merrell and Henderson (1997), which, when applied to a practical example, suggests that effect sizes of about r = 0.15 to 0.20 (small to medium, according to Cohen's standards) would equal the learning gain in one school year, which they consider an effect of huge practical relevance. Seen in this light the effect size of r = 0.11 that we arrive at when we average the results summarized in Table 3.4, might perhaps be upgraded in its rating for practical significance. Yet, the literature on estimating year effects of schooling, shows important differences between subjects, grade levels and national contexts; with coefficients as high as 0.45 (Luyten, 2007), this yardstick against which to compare leadership effects is not a very stable one.

Given the long causal chain between leadership actions and student achievement results, small effect sizes should not really come as a surprise with the kind of research designs that were used in the majority of studies analyzed. In fact it is rather the effect sizes in the order of magnitude of r = 0.40 that should be seen as remarkable.

Method

Literature Search

To identify potential relevant studies the following online databases were used: Web of science (www.isiknowledge.com), Scopus (www.scopus.com), ERIC and Psycinfo (provided through Ebscohost). The search was carried out in November 2010 and focused on publications between 2005 and 2010. The databases were searched using the key terms that were also used in the meta-analyses published in Scheerens et al. (2007).

The databases were searched using a combination of the following groups of key terms:

[4] According to Cohen (1998), small effects are in the order of r = 0.10, medium effects r = 0.30 and large effects r = 0.50 or higher.

- ‘school effectiveness’, ‘education* effectiveness’, ‘teach* effectiveness’, ‘effective* teaching’, ‘effective instruction’, ‘instruction* effectiveness’, ‘mastery learning’, ‘constructivist teaching’, ‘mathematics instruction’, ‘reading instruction’, ‘science instruction’, ‘classrooms’, ‘mathematics teaching’, ‘reading teaching’, ‘science teaching’;
- ‘value added’, attainment, achievement, ‘learn* result*’, ‘learn* outcome*’, ‘learn* gain’, ‘student* progress’;
- leadership, principal.

In total 303 hits were found (see Table 3.4). After removing the duplicate publications 255 unique publications were left.

Table 3.4
Results literature search

Database	**Number of hits**
ERIC	37
PsycInfo	140
Scopus	84
Web of Science	42
Total	**303**
Duplicates	48
Total number of possible relevant publications	**255**

In addition to the search in databases volumes of the following journals were searched:

- *American Educational Research Journal*
- *Educational Administration Quarterly*
- *Educational Management Administration & Leadership*
- *International Journal of Leadership in Education*
- *Journal of Educational Administration*
- *Leadership and Policy in Schools*
- *School Leadership and Management*
- *School Effectiveness and School Improvement*

Finally, recent reviews and books on school leadership and school effectiveness, as well as references in recent articles were checked in order to find additional literature.

Selection of Studies for the Meta-Analysis

The first selection of the studies collected was guided by the following selection criteria:

1. *Independent variable:* The study is designed explicitly to examine school leadership.
2. *Dependent variable:* The study had to include an explicit measure of cognitive student achievement.

3. *Language of the publication*: Publications included had to be written in English or Dutch. Databases and journals other than primarily English were not searched.
4. *Study population*: The study had to be conducted at primary and/or secondary school level (for students aged 4-18).
5. *Year of publication:* The study is published or presented not earlier than January 2005 and before January 2011.
6. *Methods:* Studies had to contain empirical data and outcomes.

Titles and abstracts of publications were evaluated on the six selection criteria. Using the above-mentioned selection criteria 80 publications remained for further evaluation.

Each of these publications was examined in full. After this second round, 25 publications were selected for meta-analysis. Of these 25 studies that met the selection criteria 10 used direct- and 15 used indirect-effect models. This article only discusses the analyses of the15 publications published between 2005 and 2010 which used indirect-effect models. Six studies examined indirect effects of leadership in primary school contexts, four in secondary schools and five studies included both primary and secondary schools.

Six studies were conducted in the US, four in Canada, two in the Flemish Community of Belgium and one each in England and the Netherlands. One study was based on data from 14 OECD countries participating in TIMSS 2007.

The 15 studies contained 34 replications, for which effect sizes were considered, where a replication represents each association of a leadership variable, intermediary variables and a specific outcome measure. So, for example a study that investigates leadership effects for reading and mathematics achievement would have two replications.

In all studies, structural equation modelling was used to examine the direct and indirect effects of leadership on achievement. In almost all studies, the design included control for student background effects, either through the use of gain scores or covariates.

In Table A1 an overview is presented of the variables used in the studies: the independent [leadership variable(s)], the antecedent and contextual variables, the intermediate variables and the dependent variable(s) used in the studies. As can be noticed, in more than half of the studies the indirect effect models include intermediate variables at more than one level.

Table A2 provides an overview of direct and indirect effects. All paths between leadership and achievement for which (in) direct effect size statistics were available or could be calculated are included in Table A2. In the last column of Table A2 each leadership variable the total effect size is presented (see also Table 3.5 for a summary).

Calculation of Effect Sizes

The number of studies and replications was considered too small to carry out a meta-analysis following the method we had employed in earlier studies, using multi-level analysis (Scheerens & Bosker, 1997; Scheerens et al. 2007), based on analysis techniques as described in Raudenbush and Bryk (1985) and Hox (2002).

For each replication the total effects were copied from the publication, whenever possible. In case these total effects were not explicitly published, they were computed from the path diagrams in the publications. The individual effect of a single path is computed by multiplying all effects included in that path. There may be (usually minor) differences between published total effects and total effects computed from the diagram, especially where the path diagrams only mention significant effects.

Non-weighted and weighted total effect sizes were calculated. The relative weight for each effect size was calculated based on the sample size, which in our case was determined by the number of schools included in each sample. In this meta-analysis sample sizes ranged from 38 to 363 schools (see Table 3.5).

Results[5]

Table 3.5 summarizes the total effects of all 34 replications, found in 15 publications. The mean magnitude of the non-weighted total effects equals *r* =0.031, which does not deviate significantly from zero, given a standard error of 0.20. The weighted summary effect is *r* = 0.048.

However, when the outlying publication from Ten Bruggencate (2009) is excluded[6], the non-weighted mean effect size would become *r* = 0.060, which deviates significantly from zero with a standard error of 0.18. The weighted mean is almost equal to the non-weighted mean: *r* = 0.061. This shows that including or excluding one publication can largely affect the conclusions, given the limited number of replications.

Total Effects

Studies in which relatively high effect sizes were found are those by Heck and Moriyama (2010); Leithwood and Jantzi (2008) and Leithwood and Mascall (2008).

[5] The complete overview of results, including direct, indirect and total effects for all replications is available in a set of tables that are published in Hendriks and Steen (2012).

[6] This study showed some highly negative effects (-0.31, -0.18 and -0.16 respectively) and on the other hand more replications (6) than all other publications in the table.

Table 3.5

Summary of total effects sizes in indirect effect studies

Author and year	Leadership measures	Achievement measure	Total effect	No. of schools	Relative weight
Day et al. (2009)	Integrated leadership (primary level)	Change in pupil outcomes over three years	.001	363	7.51
	Integrated leadership (secondary level)	Idem	.04	309	6.4
De Maeyer et al. (2007)	Integrated leadership	Reading	-.02	47	0.97
	idem	Math	-.16	47	0.97
Heck & Hallinger (2009)	Initial distributed leadership	Growth Rate Math	.03	195	4.04
	Change in leadership	idem	.09	195	4.04
Heck & Hallinger (2010)	Distributed leadership	Initial Reading scores (year 2)	.02	197	4.08
	idem	Initial Math scores (year 2)	.02	197	4.08
	Change in leadership	Growth Rate Reading	.10	197	4.08
	idem	Growth Rate Math	.10	197	4.08
Heck & Moriyama (2010)	Collaborative leadership	Added Year Effect Reading	.16	198	4.1
	idem	Added Year Effect Math	.14	198	4.1
Leithwood & Jantzi (2008)	Integrated leadership: School leadership	Proportion of students reading or exceeding the state's proficient level	.24	79	1.64
Leithwood et al. (2006)	School leadership	2 year mean achievement score	.11	88	1.82
	idem	2 year mean achievement gain	-.06	88	1.82
Leithwood & Mascall (2008)	Collective leadership	Percentage of students meeting or exceeding the proficiency level on language and math tests	.24	90	1.86
Leithwood, Patten & Jantzi (2010)	Distributed leadership	Percentage of students per school achieving level 3 or higher at math and literacy test	.11	199	4.12
	idem	idem	.15	199	4.12
Opdenakker & Van Damme (2007)	Participative professionally oriented leadership	Math	.002	57	1.18
Ross & Gray (2006)	Transformational leadership	Composite school score	.22	205	4.25

Author and year	Leadership measures	Achievement measure	Total effect	No. of schools	Relative weight
Seashore Louis, Dretzke & Wahlstrom (2010)	Instructional leadership	Percentage of students at school level meeting or exceeding the proficiency level 2005 math tests	.05	106	2.2
	Shared leadership	idem	.03	106	2.2
Supovitz, Sirinides & May (2010)	Principal leadership	English Language & Arts	.03	38	0.79
	idem	Math	-.009	38	0.79
Ten Bruggencate (2009)	Leadership style: Rational goals (teacher perceptions)	Average exam mark	-.16	97	2.01
	Leadership style: Internal Process (teacher perceptions)	idem	.003	97	2.01
	Leadership style: Human relations (teacher perceptions)	idem	.004	97	2.01
	Leadership style: Open systems (teacher perceptions)	Idem	-.18	97	2.01
	Leadership style: Rational goals (principal perceptions)	idem	.002	97	2.01
	Leadership style: Open systems (principal perceptions)	idem	-.31	97	2.01
Ten Bruggencate, Luyten & Scheerens (2010)	Time spent on instructional leadership	Math (TIMSS)	.02	Varies from 67 in Cyprus to 239 in US (average = 154)	3.18
	Time spent on administrative duties	idem	-.09	idem, average = 154	3.18
	Time spent on supervising teachers	Idem	.09	idem, average = 154	3.18
	Time spent on public relations	idem	.04	idem, average = 154	3.18
Mean (unweighted)	15 publications; 34 effect measures		**.031**		
SE mean			(.020)		
Mean (weighted)			**.048**		
without Ten Bruggencate (2009)	14 publications; 28 effect measures				
Mean (unweighted)			**.060**		
SE mean			(.018)		
Mean (weighted)			**.061**		

In almost all indirect studies the measurement of school leadership includes aspects of transformational leadership, and in half of these studies also instructional leadership. A more detailed description of the way school leadership was operationalized in these studies is provided in Hendriks and Steen (2012).

These results confirm earlier patterns of outcomes, where leadership studies conducted in North America and Australia tend to show somewhat higher effects than studies in European countries (Scheerens & Bosker 1997; Witziers, Bosker & Krüger, 2003).

Promising Paths and Intermediate Variables in Indirect Effect Models

In Table 3.6, an overview is given of the most promising paths in the indirect models reviewed in this study. The combined effects represent the product of the association of leadership with a particular intermediate variable and the association of the intermediate variable and student outcomes. Remarkable outcomes are the *negative* paths in the studies by De Maeyer et al. (2007) and Ten Bruggencate (2009). Negative associations are sometimes interpreted as compensatory action of schools and school leaders as a reaction to low student performance, but these interpretations are rather speculative given the correlational nature of the studies in question. Combined effects range from r = -0.32 - 0.25, with academic climate and instructional practices as the most promising intermediary variables, as far as the size of the combined effects is concerned.

Table 3.6
The most relevant paths found in indirect effect models

Author and year	Leadership measure	Achievement measure	Intermediate variables	Combined effect via this part
Day et al. (2009)	Integrated leadership (secondary level)	Change in pupil outcomes over three years	Leadership Distribution in the school	.04
De Maeyer et al. (2007)	Integrated leadership	Reading	none	-.27
	Integrated leadership	Reading	Academic climate	.25
	Integrated leadership	Math	none	-.15
Heck & Hallinger (2009)	Initial distributed leadership	Growth Rate Math	Change in Capacity	.03
	Change in leadership	Growth Rate Math	Change in Capacity	.08
Heck & Hallinger (2010)	Distributed leadership	Initial Reading scores (year 2)	School capacity	.02
	Distributed leadership	Initial Math scores (year 2)	School capacity	.02
	Change in leadership	Growth Rate Reading	Change in Capacity	.10
	Change in leadership	Growth Rate Math	Change in Capacity	.13
Heck & Moriyama (2010)	Collaborative leadership	Added Year Effect Reading	Instructional Practices	.16
	Collaborative leadership	Added Year Effect Math	Instructional Practices	.14
Leithwood & Jantzi (2008)	Integrated leadership: School leadership	Proportion of students reading or exceeding the state's proficient level	School conditions	.24
Leithwood et al. (2006)	School leadership	2 year mean achievement score	(not reported)	.11
	idem	2 year mean achievement gain	(not reported)	-.06
Leithwood & Mascall (2008)	Collective leadership	Percentage of students meeting or exceeding the proficiency level on language and math tests	(not reported)	.24
Leithwood et al. (2010)	Distributed leadership	Percentage of students per school achieving level 3 or higher at math test	Rational Path (model 1)	.14
			Emotional Path	.03
			Organizational Path	-.05
			Family Path	-.02

Author and year	Leadership measure	Achievement measure	Intermediate variables	Combined effect via this part
	Distributed leadership	Percentage of students per school achieving level 3 or higher at math and literacy test	Rational Path: Academic press (model 2)	.08
			Emotional Path: Collective teacher efficacy	.03
			Emotional Path: Teacher trust on others	.06
			Organizational Path: Instructional time	-.04
			Organizational Path: Professional learning community	.18
Ross & Gray (2006)	Transformational leadership	Composite school score	Teacher commitment to School Mission	.13
	Transformational leadership	Composite school score	Collective Teacher efficacy & Teacher commitment to Community Partnerships	.12
	Transformational leadership	Composite school score	Teacher commitment to Professional Community	-.07
Seashore Louis et al. (2010)	Instructional leadership	Percentage of students at school level meeting or exceeding the proficiency level 2005 math tests	Focused instruction	.03
	Instructional leadership	idem	Professional Community & Focused instruction	.02
	Shared leadership	idem	Professional Community & Focused instruction	.03
Supovitz et al. (2010)	Principal leadership	English Language & Arts	Change in Instruction	.02
Ten Bruggencate (2009)	Leadership style: Rational goals (teacher perceptions)	Average exam mark	none	-.16

Author and year	Leadership measure	Achievement measure	Intermediate variables	Combined effect via this part
	Leadership style: Open systems (teacher perceptions)	Average exam mark	none	-.18
	Leadership style: Open systems (principal perceptions)	Average exam mark	none	-.32
Ten Bruggencate et al. (2010)	Time spent on instructional leadership	Math achievement	Valuing Math	.02
	Time spent on administrative duties	Math achievement	none	-.04
	Time spent on administrative duties	Math achievement	Valuing Math	-.02
	Time spent on administrative duties	Math achievement	Topic coverage (OTL)	-.03
	Time spent on supervising teachers	Math achievement	none	.04
	Time spent on supervising teachers	Math achievement	Valuing Math	.01
	Time spent on supervising teachers	Math achievement	Topic coverage (OTL)	.03
	Time spent on public relations	Math achievement	Topic coverage (OTL)	.04

In Table 3.7, the intermediate variables in promising paths of the indirect-effect models are grouped according to the four core leadership functions, included in the conceptual introduction (Table 3.2).

Quite a few studies address several of the core functions. The overview in Table 3.6 shows that intermediate variables used in the studies by Heck and Hallinger (2009, 2010), Leithwood and Jantzi (2008) and Leithwood et al. (2010) cover a broad spectrum of effectiveness enhancing school factors.

In other studies, the intermediate variables were more focused on specific effectiveness enhancing variables. In the study by De Maeyer et al. (2007), the intermediate variable academic climate was limited to climate and values (high expectations and shared sense of purpose among teachers). In the study by Ten Bruggencate et al. (2010), the way students valued mathematics and topic coverage were used as intermediate variables.

Table 3.7

Connection between leadership emphases and intermediary conditions

Leadership emphasis (Leithwood)	**Main categories of intermediary conditions**
Setting directions	<u>Academic climate</u> Academic climate (De Maeyer et al., 2007) Teacher Commitment to the school mission (Ross & Gray, 2006)
Developing people	<u>Professional capacity of the staff, cooperation and commitment of staff</u> Change in school academic capacity (Heck & Hallinger, 2009) Change in school improvement capacity (Heck & Hallinger, 2010) School instructional practices (Heck & Moriyama, 2010) School conditions (Leithwood & Jantzi, 2008) Teacher's professional community (Seashore Louis et al., 2010)
Redesigning the organization	<u>Organizational capacity</u> Collective teacher efficacy (Ross & Gray, 2006) School conditions (Leithwood & Jantzi, 2008) Teacher commitment to the school as a professional community (Ross & Gray, 2006) Leadership Distribution in the school (Day et al., 2009)
Managing the teaching and learning program	<u>Instructional conditions</u> School instructional practices (Heck & Moriyama, 2010) Focused instruction (Seashore Louis et al., 2010) Change in instruction (Supovitz, 2008) Topic coverage (Ten Bruggencate et al., 2010) School conditions (Leithwood & Jantzi, 2008) "Rational Path", including academic press and disciplinary climate (Leithwood et al., 2010)

In the study by Seashore Louis et al. (2010), two intermediate variables were used: focused instruction and teachers' professional community: Focused instruction is targeted at aspects of indirect and constructivist teaching. Teachers' professional community is a variable in which several subcategories with regard to professional capacity of the staff (HRM and HRD) and climate are combined.

Intermediate variables covering aspects of teaching are included to a limited extent in the publications examined; only three studies (Heck & Moriyama 2010; Seashore Louis et al., 2010; Supovitz 2008) included teaching variables.

When comparing the theoretically derived intermediary conditions from Table 3.2, with those that were found in our review and are listed in the second column of Table 3.7, we can conclude that they can indeed be subsumed under the four key leadership emphases. Next, there is a fair correspondence between the conditions from school effectiveness research and the intermediary variables used in our indirect leadership effect studies.

The data on promising indirect paths to leadership effects are still too limited to draw strong conclusions about the relative importance of the intermediary variables. The results summarized in Table 3.7 suggest that each of the four sets of intermediary conditions (labelled under the headings of setting directions, developing people, developing the organization and managing the teaching and learning program) could play a role in explaining indirect school leadership effects. Yet, a more specific connection to instructional effectiveness seems to be a promising direction, as illustrated particularly in the studies by Heck and Moriyama (2010), Seashore Louis et al. (2010), and Ten Bruggencate et al. (2010). Heck and Moriyama (2010, p. 397) report that they:

> ... found support for (their) temporal ordering of leadership, instructional practices, and added-year effects, net of the context and social composition of the school. More specifically, stronger perceptions about leadership for learning (e.g. broad participation in improvement efforts, ongoing evaluation) were related to subsequent stronger views about the quality of instructional practices (i.e. teaching activities, learning environment) which in turn positively influenced added-year effects".

These outcomes match key assumptions of integrated educational effectiveness models, where conditions at school level are seen as relevant to the extent that they support and facilitate conditions at classroom level. As Heck and Moryiama conclude: 'The results provide support concerning the relevance of school leadership as a means of facilitating school improvement through building instructional practices in the school' (2010, p. 397).

Further quantitative and qualitative work would be needed to strengthen the knowledge base on indirect leadership effect models and obtain more detailed information on how the respective intermediary conditions work (and possibly interact) in influencing student achievement. A semi-structured qualitative approach, for instance, might take the available research results and operational definition of the key intermediary variables as a starting point for qualitative reflection of acting school leaders, in order to better understand the way they perceive indirect causation in their work.

Discussion, School Leadership as Meta Control

When examining the results of school leadership effect studies over almost three decades we find rather small direct and indirect leadership effects. Theoretical work on the school as an organization clarifies why we should not have expected high leadership effects in the first place. We have noted that in the development of school leadership concepts over time the notion that schools have many 'substitutes' for leadership has been re-discovered, and in some recent studies of distributed and organizational leadership, focused action of one central leader has practically disappeared from the scene. The theoretical work and results of empirical studies highlighted in our study suggest that in 'normal' situations of average schools a 'lean' kind of management might be sufficient, which would make maximum use of the available substitutes and self-organization offered by the school staff and other provisions. This kind of management fits the concept of 'meta control', which could be interpreted as orchestrating the control by the other actors on the school scene.

The concept of meta control originates from control theory (De Leeuw, 1990). According to his 'control paradigm', four major types of direct control can be distinguished: routine control, adaptive control, goal control and environmental control. Routine control is about the day-to-day monitoring of an organization's primary process. In the case of schools, under normal circumstances, very little monitoring of teachers and teaching is required. The narrow interpretation of instructional leadership is close to this kind of routine control. Meta control directed at this kind of routine control could be seen as creating favorable conditions for teachers to do their work independently. It could mean that, on the one hand, the school leader protects teachers against disturbing external influences (buffering) and, on the other hand provides facilitation in the sense of opportunities for professional development, alignment among staff, feedback, and provision of the necessary teaching resources. Taking care of and overseeing administrative and clerical tasks of the school could be seen as part of the buffering function as well. Adaptive control refers to the supervision and change of the organization's structure and core processes. Creation of new structures for teacher cooperation and the school-wide adoption of specific ICT applications are examples of adaptive control. Alignment versus loose coupling is the theoretical issue that is at stake here (cf. Elmore, 2000). Meta control directed at organizational structures and key processes is close to the management of change and transformational leadership. A school leader as a meta-controller would also need to oversee the pros and cons of structural school reform, as compared to 'simply' optimizing normal functioning (routine control). Similarly, adaptively oriented meta control would have to strike a balance between supportive, routine and innovative aspects of the functioning of the organization. Goal control has to do with upholding performance standards and soliciting agreement on the core objectives of the organization. Goal control as meta control recognizes that, within the controlled system, in our case the school's sub-units (i.e. teachers) have their own goals. In this case the meta-controller has a task in coordinating the individual goals and uniting them under a common school mission. This kind of goal control is close to the extended view of instructional leadership, and to transformational leadership. In environmental control, leaders influence the functioning of the organization by means of putting into play stimulants from the environment. This role of leadership becomes more important as schools are increasingly operating in networks or as part of higher-level organizations, such as school districts.

In short, school leaders as meta-controllers need to have a broad overview of key areas of organizational functioning, a keen eye for self-steering and self-organization and a detached attitude to taking matters in their own hand (diverting from meta-, to direct control).

References

Baumert, J., Lüdtke, O., & Trautwein, U. (2006). *Interpreting effect sizes in large-scale educational assessments.* Berlin: Max Planck Institute for Human Development.

Bolman, L. G., & Deal, T. E. (1997). *Reframing organizations: Artistry, choice, and leadership* (2nd ed.). San Francisco, CA: Jossey-Bass.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: Wiley.

Bossert, S. T., Dwyer, D. C., Rowan, B., & Lee, G. V. (1982). The instructional management role of the principal. *Educational Administration Quarterly, 18*, 34-64. doi:10.1177 /0013161X82018003004

Chin, J. M-C. (2007). Meta-analysis of transformational school leadership effects on school outcomes in Taiwan and the USA. *Asia Pacific Education Review, 8*, 166-177. doi:10.1007/BF03029253

Cohen, J. (1998). Statistical power analysis for the behavioral sciences. 2nd edition. Hillsdale, NJ: Lawrence Erlbaum

Creemers, B., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. New York, NY: Routledge.

De Leeuw, A. C. J. (1990). *Organisaties: management, analyse, ontwerp en verandering: Een systeem visie* [Organisations: management, analysis, design and change: A system's view]. Assen/Maastricht: Van Gorcum.

Edmonds, R. (1979). Effective schools for the urban poor. *Educational Leadership,* 37, 15-24. Retrieved from: http://www.ascd.org/ASCD/pdf/journals/ed_lead/el_197910_edmonds .pdf

Elmore, R. F. (2000). *Building a new structure for school leadership*. Washington, DC: The Albert Shanker Institute.

Glatthorn, A. A. (1987). *Curriculum leadership*. London: Glenview.

Glatthorn, A. A. (1997). *The principal as curriculum leader*. Thousand Oaks, CA: Corwin Press.

Hallinger, P. (1983). *Assessing the instructional management behavior of principals* (Unpublished doctoral dissertation). Stanford, CA: Stanford University. ERIC Document No. 8320806.

Hallinger, P., & Heck, R. H. (2009). Distributed leadership in schools: Does system policy make a difference? In A. Harris (Ed.), *Distributed leadership: Different perspectives*, *7* (pp. 101-117). London: Springer Science+Business Media B.V. doi:10.1007/978-1-4020-9737-9

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

Heck, R. H., & Moriyama, K. (2010). Examining relationships among elementary schools' contexts, leadership, instructional practices, and added-year outcomes: A regression discontinuity approach. *School Effectiveness and School Improvement,* 21, 377-408. doi:10.1080/09243453.2010.500097

Hendriks, M. A., & Steen, R. (2012). Results from school leadership effectiveness studies (2005-2010). In J. Scheerens (Ed.)*, School leadership effects revisited. Review and meta-analysis of empirical studies* (pp. 65-129). Dordrecht: Springer Research Briefs in Education.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Lawrence Erlbaum Associates.

Leithwood, K. (1992). Transformational leadership: The move toward transformational leadership. *Educational Leadership,* 49(5), 8-12. Retrieved from: http://www.ascd.org /ASCD/pdf/journals/ed_lead/el_199202_leithwood.pdf

Leithwood, K., Jantzi, D., & McElheron-Hopkins, Ch. (2006). The development and testing of a school improvement model. *School Effectiveness and School Improvement*, 17, 441-464. doi:20.1080/09243450600743533

Luyten, H. (2007). *Opportunities to utilize already existing quantitative data for PISA 2009.* (Discussion paper). Enschede: University of Twente.

Marks, H. M., & Printy, S.M. (2003). Principal leadership and school performance: An integration of transformational and instructional leadership. *Educational Administration Quarterly, 39*, 370-397. doi:10.1177/0013161X03253412

Marzano, R. J., Waters, T., & McNulty, B. A. (2005). *School leadership that works: From research to results*. Alexandria, VA: Association for Supervision and Curriculum Development.

Mulford, B. (2003). *School leaders: Changing roles and impact on teacher and school effectiveness.* Paper commissioned by the Education and Training Policy Division for the activity Attracting, Developing and Retaining Effective Teachers. Paris: OECD. Retrieved from: http://www.oecd.org/dataoecd/61/61/2635399.pdf

Quinn, R. E., Faerman, S. R., Thompson, M. P., & McGrath, M. R. (1996). *Becoming a master manager: A competency framework*. New York: John Wiley.

Raudenbush, S., & Bryk, A. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10*, 75-98. doi:10.2307/1164836

Richard, F. D., Bond, C. F. Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331-363. doi:10.1037/1089-2680.7.4.331

Robinson, V. M. J., Lloyd, C., & Rowe, K. J. (2008). The impact of leadership on student outcomes: An analysis of the differential effects of leadership types. *Educational Administration Quarterly, 44*, 635-674. doi:10.1177/0013161X08321509

Scheerens, J. (2000). *Improving school effectiveness*. Fundamentals of Educational Planning series no. 68. Paris: UNESCO, International Institute for Educational Planning.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.

Scheerens, J., Luyten, H., Steen R., & Luyten-de Thouars, Y. (2007). *Review and meta-analyses of school and teaching effectiveness*. Enschede: University of Twente.

Silins, H., & Mulford, B. (2004). Schools as learning organisations. Effects on teacher leadership and student outcomes. *School Effectiveness and School Improvement*, 443-466. doi:10.1080/09243450512331383272

Supovitz, J. A. (2008). Instructional leadership in American high schools. In M. M. Mangin & S. R. Stoelinga (Eds.), *Effective teacher leadership: Using research to inform and reform*, (pp. 144-162). New York: Teachers College Press.

Ten Bruggencate, G., Luyten, H., & Scheerens, J. (2010). *Quantitative analysis of international data, exploring indirect effect models of school leadership*. Enschede: University of Twente.

Tymms, P., Merrell, C., & Henderson, B. (1997). The first year at school. A quantitative investigation of the attainment and progress of pupils. *Educational Research and Evaluation, 3*, 101-118. doi:10.1080/1380361970030201

Witziers, B., Bosker, R. J., & Krüger. M. L. (2003). Educational leadership and student achievement: The elusive search for an association. *Educational Administrative Quarterly* 39, 398-425. doi:10.1177/0013161X03253411

Indirect effect studies used for meta-analyses

Day, C., Sammons, P., Hopkins, D., Harris, A., Leithwood, K., Gu, Q., Brown, E., Ahtaridou, E., & Kington, A. (2009). *The impcat of school leadership on pupil outcomes*. Nottingham, UK: The National College for School Leadership.

De Maeyer, S., Rymenans, R., Petegem, P. van, Bergh, H. van den, & Rijlaarsdam, G. (2007). Educational leadership and pupil achievement: The choice of a valid conceptual model to test effects in school effectiveness research. *School effectiveness and School Improvement, 18*, 125-145. doi:1080/09243450600853415

Heck, R. H., & Hallinger, Ph. (2009). Assessing the contribution of distributed leadership to school improvement and growth in math achievement. *American Educational Research Journal, 46,* 659-689. doi:10.3102/0002831209340042

Heck, R. H., & Hallinger, Ph. (2010). Testing a longitudinal model of distributed leadership effects on school improvement. *The Leadership Quarterly, 21,* 867–885. doi:10.1016/j.leaqua.2010.07.013

Heck, R. H., & Moriyama, K. (2010). Examining relationships among elementary schools' contexts, leadership, instructional practices, and added-year outcomes: a regression discontinuity approach. *School Effectiveness and School Improvement, 21*, 377-408. doi:10.1080/09243453.2010.500097

Leithwood, K., & Jantzi, D. (2008). Linking leadership to student learning: The contributions of leader efficacy. *Educational Administration Quarterly, 44,* 496-528. doi:10.1177/0013161X08321501

Leithwood, K., Jantzi, D., & McElheron-Hopkins, Ch. (2006). The development and testing of a school improvement model. *School Effectiveness and School Improvement, 17*, 441-464. doi:10.1080/09243450600743533

Leithwood, K., & Mascall, B. (2008). Collective leadership effects on student achievement. *Educational Administration Quarterly, 44,* 529-561. doi:10.1177/0013161X08321221

Leithwood, K, Patten, S., & Jantzi, D. (2010). Testing a conception of how school leadership influences student learning. *Educational Administration Quarterly 46,* 671-706. doi:10.1177/0013161X10377347

Opdenakker, M-C., & 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. doi:10.1080/01411920701208233

Ross, J. A., & Gray, P. (2006). School leadership and student achievement. The mediating effects of teacher beliefs. *Canadian Journal of Education, 29*, 798-822. doi:10.2307/20054196

Seashore Louis, K., Dretzke, B., & Wahlstrom, K. (2010). How does leadership affect student achievement? Results from a national US survey. *School Effectiveness and School Improvement, 21*, 315 -336. doi:10.1080/09243453.2010.486586

Supovitz, J., Sirinides, Ph., & May, H. (2010). How principals and peers influence teaching and learning. *Educational Administration Quarterly 46*, 31-56. doi:10.1177 /1094670509353043

Ten Bruggencate, G.C. (2009). *Maken schoolleiders het verschil?* [Do school leaders make a difference?]. (Unpublished doctoral dissertation). Retrieved from: http://dx.doi.org /10.3990/1.9789036527835

Ten Bruggencate, G., Luyten, H., & Scheerens., J. (2010). *Quantitative analysis of international data, exploring indrect effect models of school leadership*. Enschede: University of Twente.

Table A1

Characteristics of indirect effect studies on school leadership

Author and Year	Country	School type	No of schools	Leadership variable	Antecedents	Intermediate level 1	Intermediate level 2	Intermediate level 3	Achievement measure
Day et al. (2010)	England	Primary	363	Integrated leadership	SES	Leadership distribution: • 'Distributed leadership • Staff • SMT[7] • SLT[8] Collaboration' • SLT's Impact on Learning and Teaching'	School processes: • Teacher collaborative culture • Assessment for learning • 'Improvement in school conditions' • External Collaborations & Learning Opportunities'	• High academic standards • Pupil motivation' and responsibility for learning • Reduction in staff mobility and absence • Change in pupil behavior • Change in pupil attendance	Change in pupil outcomes over three years
Day et al. (2010)	England	Secondary	309	Integrated leadership	SES	Leadership distribution: • Distributed leadership • Staff' • SLT Collaboration • SLT impact on learning and teaching	School processes: • Teacher collaborative culture • Assessment for Learning • Improvement in school conditions • External collaborations and learning opportunities'	• High academic standards • Pupil motivation and learning culture • Change in pupil behavior • 'Change in pupil attendance.	Change in pupil outcomes over three years
Heck & Hallinger (2009)	US	Primary	195	Initial distributed leadership		Change in Capacity	School Organization		Math Growth

[7] SMT: School Mangement Team

[8] SLT: Senior Leadership Team

Author and Year	Country	School type	No of schools	Leadership variable	Antecedents	Intermediate level 1	Intermediate level 2	Intermediate level 3	Achievement measure
				(three aspects) • School improvement • School governance • Resource management					
				Initial distributed leadership		Change in Capacity			Math Growth
				Change in leadership	Principal stability	Change in Capacity	School Organization		Math Growth
				Change in leadership	Principal stability	Change in Capacity			Math Growth
Heck & Hallinger (2010a) Leadership Quarterly	US	Primary	197	Transformational leadership Distributed leadership (three aspects) • School improvement • School governance • Resource management		School improvement capacity.			Math (grade 3-5) Reading (grade 3-5)
				Initial Leadership (year 1)	Student composition	Initial School Capacity (year 1)			Initial Read level (year 2)
				Initial Leadership (year 1)	Student composition	Initial School Capacity (year 1)			Initial Math level (year 2)
				Change in Leadership	Teacher stability, principal stability school size	Change in school capacity			Growth Rate Read
				Change in Leadership		Change in school capacity			Growth Rate Math

Author and Year	Country	School type	No of schools	Leadership variable	Antecedents	Intermediate level 1	Intermediate level 2	Intermediate level 3	Achievement measure
Heck & Moriyama (2010)	US	Primary	198	Collaborative leadership: shared school governance, collaborative decisions focusing on academic improvement, broad participation in efforts to evaluate the school's academic development	Student stability Student composition Teacher stability and experience	Instructional practices (four dimensions); Classroom teaching conditions, Quality of student support system Professional capacity Focused and sustained action on improvement			Added year effect Read
				Collaborative leadership		Instructional practices			Added year effect Math
Leithwood & Jantzi (2008)	US	Primary and Secondary (K-12)		School leadership: 1. Setting directions 2. Developing people 3. Redesigning the organization 4. Managing the instructional program	District leadership • District conditions • Leaders' Self Efficacy Beliefs • Leaders' Collective Efficacy Beliefs Co-variables: District size School size School level Principal turnover	School conditions • School culture • Decision making processes • Supports for instruction • Professional learning community	Class conditions • Workload • Areas of formal preparation • Student grouping • Curriculum and instruction		Percentages of students meeting or exceeding the proficiency level on language and math tests (averaged across grades and subjects - language and mathematics) over 3 years (2003 to 2005)

Author and Year	Country	School type	No of schools	Leadership variable	Antecedents	Intermediate level 1	Intermediate level 2	Intermediate level 3	Achievement measure
					Principal gender, Years of experience, Race, Ethnicity				
Leithwood, Jantzi & McElheron-Hopkins (2006)	Canada	Primary	88	Transformational school leadership • providing resources and support, empowering teams • facilitative: encourage successful implementation		School Improvement Planning Process	Contents of the School Improvement Plan Implementation process		Students' two year mean achievement scores (as measured by the provincial tests of literacy and mathematics in Grades 3 and 6) Students' two year mean gain
Leithwood & Mascall (2008)	Canada	Primary and secondary	90	Transformational leadership: Collective leadership	SES	Capacity Motivation Setting			Percentage of students meeting or exceeding the proficiency level on language and math tests (averaged across grades and subjects - language and mathematics) over 3 years (2003 to 2005)
Leithwood, Patten & Jantzi (2010)	Canada	Primary	199	Distributed leadership (with focus on managing and leading the instructional program	Composite SES	Rational Path Emotional Path Organizational Path Family Path			Percentage of students per school achieving level 3 or higher at the grade 3 and 6 math and literacy achievement

Author and Year	Country	School type	No of schools	Leadership variable	Antecedents	Intermediate level 1	Intermediate level 2	Intermediate level 3	Achievement measure
						Rational Path: • Academic Press • Disciplinary Climate Emotional Path: • Collective Teacher Efficacy • Teacher Trust in Others Organizational Path • Instructional Time • Professional Learning Community Family Path: • Computer at Home • Adult Help at Home			
Louis et al. (2010)	US	Primary and Secondary	106 (50 primary 53 secondary, 3 K-8)	Instructional leadership Shared leadership	School poverty	Professional community	Focused instruction		Percentages of students at school level meeting or exceeding the proficiency level on 2005 math tests
De Mayer et al. (2007)	Belgium (Flanders)	Secondary	47	Integrated leadership	Percentage of girls, Mean IQ, Mean SES, Mean linguistic ethnic background (LEB)	Academic climate			Mathematics Reading

Author and Year	Country	School type	No of schools	Leadership variable	Antecedents	Intermediate level 1	Intermediate level 2	Intermediate level 3	Achievement measure
Opdenakker & Van Damme (2007)	Belgium (Flanders)	Secondary	57	Integrated leadership Participative professionally oriented leadership	School size Average intellectual level of students in school	Cooperation between teachers School climate: • Learning climate • Relational climate Opportunity to learn			Mathematics test
Ross & Gray (2006)	Canada	Primary	205	Transformational Leadership	SES	Collective Teacher Efficacy	Teacher Commitment to School Mission Teacher Commitment to Professional Community Teacher Commitment to Community Partnerships		Grade 3 and 6 Achievement: Residuals from regression 2001 scores over 2000 averaged across grades (3 and 6) en subjects (reading, writing and mathematics)
Supovitz et al. (2010)	US	Primary and secondary	52	Principal Leadership		Peer Influence	Change in Instruction		English Language Arts Mathematics
Ten Bruggencate (2009)	Nether-lands	Secondary		School leadership style Rational goals Internal Process Human Relations Open systems	School size Denomination (RC) Percentage of cultural minorities Urbanization of the environment Valuing education Competition	Performance-orientation Development-rientation	Student engagement Teacher's work		Promotion rate Average exam mark

Author and Year	Country	School type	No of schools	Leadership variable	Antecedents	Intermediate level 1	Intermediate level 2	Intermediate level 3	Achievement measure
Ten Bruggencate et al. (2010)	14 countries (TIMSS 2007)	Secondary	Varies from 67 in Cyprus to 239 in US[9]	Time spent on Instructional leadership Administrative duties Supervising teachers Public relations	No. of books at home; Total school enrollment; Type of community; Disadvantaged students limiting teaching; Student behavior frequency; Shortage of ICT resources limiting teaching; School climate (principal's perception)	Student attitudes towards mathematics (valuing math) Student attitudes towards school (valuing school)			Math achievement

[9] In order to correct for over and underrepresentation of schools a weighting procedure was applied to ensure that within countries all schools are represented appropriately. The weighting procedure also ensures that all countries have the same weight in the overall analyses regardless of their size.

Table A2

Direct and indirect paths from school leadership to student achievement and total effect sizes

Author and Year	Leadership measure	Achievement measure	Direct effect ES*[10]	Indirect effect level 1		Indirect effect level 2		Indirect effect level 3	
				Label	ES*	Label	ES*	Label	ES*
Day et al. (2010)	Integrated leadership: (primary level)	Change in pupil outcomes over three years		Leadership distribution in the school	.28	Teacher Collaborative Culture	.13	Pupil Motivation & Responsibility for Learning	.21
Day et al. (2010)[11]	Integrated leadership: (secondary level)	Change in pupil outcomes over three years		Leader-ship distribution in the school	.29	Change in pupil outcomes	.12		
				Contribution of all other paths with 3 to 6 intermediate levels	.005				
	Total								
Heck & Hallinger (2009)	Initial Distributed leadership	Math Growth Rate		Change in Capacity	.14**	School Organization	.22**	Math Growth Rate	.09**
	Initial Distributed leadership	Math Growth Rate		Change in Capacity	14**	Math Growth Rate	.18**		
	Total								
	Change in Leadership	Math Growth Rate		Change in Capacity	.46**	School Organization	.22**	Math Growth Rate	.09**
	Change in Leadership	Math Growth Rate		Change in Capacity	46**	Math Growth Rate	.18**		
	Total								
Heck & Hallinger (2010a) Leadership Quarterly	Transformational leadership	Initial Reading scores (year 2)		School capacity	.12**	Initial Read level (year 2)	.13**		
	Distributed leadership								
	Transformational	Initial Math scores		School capacity	.12**	Initial Math level	.15**		

[10] (*p< 0.10, ** p<.05, *** p<0.01)

[11] The model is too complex to describe in full. Only the most important path and the contribution of all other paths together are included.

Author and Year	Leadership measure	Achievement measure	Direct effect	Indirect effect level 1		Indirect effect level 2		Indirect effect level 3	
			ES*[10]	Label	ES*	Label	ES*	Label	ES*
	leadership	(year 2)				(year 2)			
	Distributed leadership								
	Transformational leadership Change in leadership	Growth Rate Reading		Change in school capacity	.49**	Growth Rate Read	.20**		
	Transformational leadership Change in leadership	Growth Rate Math		Change in school capacity	.49**	Growth Rate Math	.20**		
Heck & Moriyama (2010)	Collaborative leadership	Added Year Effect Read		Instructional Practices	.26**	Added Year Effect Read	.57**		
	Collaborative leadership	Added Year Effect Math		Instructional Practices		Added Year Effect Math	.49**		
Leithwood & Jantzi (2008)	Integrated leadership School leadership	Proportion of students reaching or exceeding the state's proficient level		School conditions	.66***	Proportion students exceeding proficient level	.40***		
Leithwood, Jantzi & McElheron-Hopkins (2006)	School leadership[12]	2 year mean achievement score							
	School leadership[13]	2 year achievement							

[12] Only the standardized total effects for independent and mediating variables on mean student achievement are presented, so it is unknown if the effects from leadership on achievement are direct or indirect

[13] Only the standardized total effects for independent and mediating variables on mean student achievement are presented, so it is unknown if the effects from leadership on achievement are direct or indirect

Author and Year	Leadership measure	Achievement measure	Direct effect ES*[10]	Indirect effect level 1 Label	ES*	Indirect effect level 2 Label	ES*	Indirect effect level 3 Label	ES*
		gain							
Leithwood & Mascall (2008)	Collective Leadership[14]	Percentage of students meeting or exceeding the proficiency level on language and math tests							
Leithwood, Patten & Jantzi (2010) Model 1	Distributed leadership	Percentage of students per school achieving level 3 or higher at math test		Rational Path	.56**	Percentage of students achieving level 3 or higher at math and literacy test	.26**		
				Emotional Path	.15**		.21		
				Organizational Path	.57**		-.08		
				Family Path	-.07		.26**		
	Total								
Leithwood, Patten & Jantzi (2010) Model 2	Distributed leadership	Percentage of students per school achieving level 3 or higher at math and literacy test		Rational Path: Academic Press	.33**	Percentage of students achieving level 3 or higher at math and literacy test	.23**		
				Rational Path: Disciplinary Climate			.22**		

[14] In the article just the total indirect effect is presented

Author and Year	Leadership measure	Achievement measure	Direct effect	Indirect effect level 1		Indirect effect level 2		Indirect effect level 3	
			ES*[10]	Label	ES*	Label	ES*	Label	ES*
				Emotional Path: Collective Teacher Efficacy	.10**		.34**		
				Emotional Path: Teacher Trust in Others	.28**		.20		
				Organizational Path : Instructional Time	.30**		-.12		
				Organizational Path : Professional Learning Community	.69**		.26**		
				Family Path: Computer at Home			-.16**		
				Family Path: Adult Help at Home					
	Distributed Leadership (total)								
Louis et al (2010)	Instructional leadership	Percentages of students at school (building) level meeting or exceeding the proficiency level 2005 math tests		Professional Community	.267***	Focused instruction	.395***	Mean math achievement	.205**
						Focused instruction	.148	Mean math achievement	.205**
	Total								

Author and Year	Leadership measure	Achievement measure	Direct effect ES*[10]	Indirect effect level 1		Indirect effect level 2		Indirect effect level 3	
				Label	ES*	Label	ES*	Label	ES*
	Shared leadership			Professional Community	.381***	Focused instruction	.395***	Mean math achievement	.205**
	Total								
Mayer et al. (2007)	Integrated leadership	Reading	-.27						
				Academic climate	.59**	Reading	.43**		
	Total								
	Integrated leadership	Mathematics	-.15						
				Academic climate	.59**	Mathematics	-.02**		
	Total								
Opdenakker & Van Damme (2007)	Participative professionally oriented leadership	Math		Cooperation	.01	Relations Climate	.75***	Learning climate	1.00**
				Cooperation	.01	Relations Climate	.75***	Learning climate	1.00**
	Total								
Ross & Gray (2006)	Transformational Leadership	Composite school score: Residuals from regression 2001 scores over 2000 scores (averaged across grades en subjects).		Collective Teacher efficacy	.48***	Teacher Commitment to School Mission	.20***	Composite school score (achievement)	.17
				Teacher Commitment to School Mission	.75***	Composite school score	.17		

Author and Year	Leadership measure	Achievement measure	Direct effect	Indirect effect level 1		Indirect effect level 2		Indirect effect level 3	
			ES*[10]	Label	ES*	Label	ES*	Label	ES*
				Collective Teacher efficacy	.48***	Teacher Commitment to Professional Community	.27***	Composite school score	-.14
				Teacher Commitment to Professional Community	.49***	Composite school score	-.14	Teacher Commitment to Professional Community	
				Collective Teacher efficacy	.48***	Teacher Commitment to Community Partnerships	.78***	Composite school score	.33***
				Teacher Commitment to Community Partnerships	.12**	Composite school score	.33***		
	Total								
Supovitz et al. (2010)	Principal Leadership	English Language Arts		Peer Influence	.38***	Change in Instruction	.21***	English Language Arts	.11**
				Change in Instruction	.18***	English Language Arts	.11**		
	Total								
	Principal Leadership	Mathematics		Peer Influence	.30***	Change in Instruction	.26***	Mathematics	-.04
				Change in Instruction	.14***	Mathematics	-.04		
	Total								

Author and Year	Leadership measure	Achievement measure	Direct effect	Indirect effect level 1		Indirect effect level 2		Indirect effect level 3	
			ES*[10]	Label	ES*	Label	ES*	Label	ES*
Ten Bruggencate (2009)	Leadership style Rational goals (teacher perceptions)	Average exam mark	-0.17**	Development orientation	0.57**	Teacher’s work	0.24**	Promotion rate	0.18**
	Total								
	Leadership style Internal Process (teacher perceptions)			Development orientation	0.23**	Teacher’s work	0.24**	Promotion rate	0.18**
	Total								
	Leadership style Human Relations (teacher perceptions)			Development orientation	0.35**	Teacher’s work	0.24**	Promotion rate	0.18**
	Total								
	Leadership style Open systems (teacher perceptions)		-.18**	Development orientation	.56**	Teacher’s work	.24**	Promotion rate	.18**
	Total								
Ten Bruggencate (2009)	Leadership style Rational goals (principal perceptions)			Development orientation	.28**	Teacher’s work	.24**	Promotion rate	.17**
	Leadership style Open systems (principal perceptions)		-.32**	Development orientation	.33*.	Teacher’s work	.24**	Promotion rate	.17**
	Total								

Author and Year	Leadership measure	Achievement measure	Direct effect	Indirect effect level 1		Indirect effect level 2		Indirect effect level 3	
			ES*[10]	Label	ES*	Label	ES*	Label	ES*
Ten Bruggencate et al. (2010)	Time spent on instructional leadership	Math achievement		Valuing math	-.12	Math achievement	-.21		
	Time spent on administrative duties		-.04						
				Valuing math	.08	Math achievement	-.21		
				Topic coverage	-.13	Math achievement	.24		
	Total								
	Time spent on supervising teachers		.04						
				Valuing math	-.06	Math achievement	-.21		
				Topic coverage	.12	Math achievement	.24		
	Total								
	Time spent on public relations			Topic coverage	.15	Math achievement	.24		

Table A2 (continued)

Direct and indirect paths from school leadership to student achievement and total effect sizes

Author and Year	Leadership measure	Achievement measure	Indirect effect level 4		Indirect effect level 5		Total effect
			Label	ES*	Label	ES*	ES*
Day et al. (2010)	Integrated leadership: (primary level)	Change in pupil outcomes over three years	Change in pupil outcomes	.18			.001
Day et al. (2010)[15]	Integrated leadership: (secondary level)	Change in pupil outcomes over three years					
	Total						.04
Heck & Hallinger (2009)	Initial Distributed leadership	Math Growth Rate					
	Initial Distributed leadership	Math Growth Rate					
	Total						0.28
	Change in Leadership	Math Growth Rate					
	Change in Leadership	Math Growth Rate					
	Total						.092
Heck & Hallinger (2010a)	Transformational leadership	Initial Reading scores (year 2)					**.02****
Leadership Quarterly	Distributed leadership						
	Transformational leadership	Initial Math scores (year 2)					**.02****
	Distributed leadership						
	Transformational leadership	Growth Rate Reading					**.10****
	Change in leadership						
	Transformational leadership	Growth Rate Math					**.13****
	Change in leadership						

[15] The model is too complex to describe in full. Only the most important path and the contribution of all other paths together are included.

Author and Year	Leadership measure	Achievement measure	Indirect effect level 4		Indirect effect level 5		Total effect
			Label	ES*	Label	ES*	ES*
Heck & Moriyama (2010)	Collaborative leadership	Added Year Effect Read					**.16****
	Collaborative leadership	Added Year Effect Math					**.14****
Leithwood & Jantzi (2008)	Integrated leadership School leadership	Proportion of students reaching or exceeding the state's proficient level					**.24*****
Leithwood, Jantzi & McElheron-Hopkins (2006)	School leadership[16]	2 year mean achievement score					.11
	School leadership[17]	2 year achievement gain					-.06
Leithwood & Mascall (2008)	Collective Leadership[18]	Percentage of students meeting or exceeding the proficiency level on language and math tests					**.24*****
Leithwood, Patten & Jantzi (2010) Model 1	Distributed leadership	Percentage of students per school achieving level 3 or higher at math test					
	Total						**.11****
Leithwood, Patten &	Distributed leadership	Percentage of students per school					

[16] Only the standardized total effects for independent and mediating variables on mean student achievement are presented, so it is unknown if the effects from leadership on achievement are direct or indirect

[17] Only the standardized total effects for independent and mediating variables on mean student achievement are presented, so it is unknown if the effects from leadership on achievement are direct or indirect

[18] In the article just the total indirect effect is presented

Author and Year	Leadership measure	Achievement measure	Indirect effect level 4		Indirect effect level 5		Total effect
			Label	ES*	Label	ES*	ES*
Jantzi (2010) Model 2		achieving level 3 or higher at math and literacy test					
	Distributed Leadership (total)						**.15****
Louis et al (2010).	Instructional leadership	Percentages of students at school (building) level meeting or exceeding the proficiency level 2005 math tests					
	Total						.052
	Shared leadership						
	Total						.031
Mayer et al. (2007)	Integrated leadership	Reading					
	Total						**-.02**
	Integrated leadership	Mathematics					
	Total						**-.16**
Opdenakker & Van Damme (2007)	Participative professionally oriented leadership	Math	Opportunity to learn	.08	Math	.14*	
			Effort	.75***	Math	.38**	
	Total						**.006**
Ross & Gray (2006)	Transformational Leadership	Composite school score: Residuals from regression 2001 scores over 2000 scores (averaged across grades en subjects).					

Author and Year	Leadership measure	Achievement measure	Indirect effect level 4		Indirect effect level 5		Total effect
			Label	ES*	Label	ES*	ES*
	Total						.220
Supovitz et al. (2010)	Principal Leadership	English Language Arts					
	Total						**.03****
	Principal Leadership	Mathematics					
	Total						-.01
Ten Bruggencate (2009)	Leadership style Rational goals (teacher perceptions)	Average exam mark					
			Average exam mark	.27**			
	Total						**-.16****
	Leadership style Internal Process (teacher perceptions)		Average exam mark	.27**			
	Total						.00
	Leadership style Human Relations (teacher perceptions)		Average exam mark	.27**			
	Total						.00
	Leadership style Open systems (teacher perceptions)						
			Average exam mark	.27**			
	Total						**-.18****
Ten Bruggencate (2009)	Leadership style Rational goals		Average exam mark	.20*			.00

Author and Year	Leadership measure	Achievement measure	Indirect effect level 4		Indirect effect level 5		Total effect
			Label	ES*	Label	ES*	ES*
	(principal perceptions) Leadership style Open systems (principal perceptions)		Average exam mark	.20*			
	Total						**-.31****
Ten Bruggencate et al. (2010)	Time spent on instructional leadership	Math achievement					.02
	Time spent on administrative duties						
	Total						-.09
	Time spent on supervising teachers						
	Total						.09
	Time spent on public relations						.04

4

Effects of evaluation and assessment on student achievement; A review and meta-analysis[1]

[1] This chapter is based on: Hendriks, M.A., Scheerens, J. & Sleegers, P.J.C., Effects of evaluation and assessment on student achievement: A review and meta-analysis (in preparation)

Abstract

In this review study and meta-analysis the evidence on the impact of evaluation and assessment as effectiveness enhancing school and classroom level conditions is summarized and updated. The meta-analysis included 20 studies on evaluation and 6 studies examining the impact of assessment. A vote count procedure was applied as well to permit the inclusion of studies that did not provide sufficient information to calculate an effect size. Findings demonstrated statistically significant but small positive effects for evaluation at school and evaluation at class level, while the average effect size for assessment was almost zero. Results of the vote count were in the same direction. The results of the conceptual analysis showed that a thorough and complete application of the evaluative cycle was rarely addressed in any of the studies included in this review. More specifically, hardly any empirical research was found on the processes by which teachers and school leaders noticed and interpreted data. A further need is also to understand the types of professional development and support that enhance effective evaluation and assessment practices.

Introduction

One of the five factors Edmonds drew forward on the basis of school effectiveness research was frequent evaluation and assessment of student performance. So from the early days of effective schools' research onwards evaluation and assessment has been mentioned as part of limited set of effectiveness enhancing conditions (Edmonds, 1979). This has not changed, and evaluation and assessment remain prominently present in recent reviews of the literature (Reynolds, Sammons, De Fraine, Van Damme, Townsend, Teddlie & Stringfield, 2014). Developments in education like structured approaches to teaching, such as mastery learning, school improvement strategies like school based review, and, at the above school level, accountability policies have further strengthened the interest in evaluation and assessment. Evaluation and assessment are increasingly considered as potential levers of change that could assist with decision-making and continuous improvement at all levels of the education system (OECD, 2013; Parr & Timperley, 2008). While evaluation and assessment traditionally focused on the assessment of students, performance data are increasingly complemented by a wide range of other data including e.g. data on student characteristics and school and instructional processes. Furthermore developments in educational measurement and psychometric theory contributed to flexibility of application and enhanced credibility (e.g. Van der Linden, 1995).

In this chapter the basic issue is the effect evaluation and assessment have on student achievement. Quite in line with the effective schools' tradition evaluation and assessment are seen as effectiveness enhancing conditions. The focus will be on the impact of evaluation and assessment at school and classroom level.

Evaluation and assessment have a place in rational planning models, like the plan, do, check, act cycle. The particular point in placing evaluation as the starting point of such cycles is that a retroactive[1], learning from experience type of approach is followed (cf., Scheerens, Glas and Thomas, 2003, p.82). Outcomes of evaluation and assessment provide focus and direction for remedial and improvement oriented action, and appeal at the same time on the achievement motivation of pupils and teachers. The way the information is fed back to the main actors is very important. When the feedback is received and registered by the main actors, it may give rise to a new planning/evaluation and feedback cycle. Briefly summarized, evaluation and assessment affect student achievement by providing substantive focus and normative attainment targets, appealing on achievement motivation and by stimulating learning on the basis of appropriate feedback.

Evaluation and assessment in schools, at school organizational and classroom level has quite a few different emphases and orientations, which will be reviewed in a section on construct analysis. Next, earlier meta-analyses of the effect of evaluation and assessment on educational achievement will be reviewed. After these introductory sections the core of this chapter describes the methods and results of a meta-analysis on the effects of evaluation and assessment on students' cognitive achievement.

Conceptualization of Evaluation and Assessment as Factors in School and Instructional Effectiveness Research

All forms of *evaluation* consist of systematic information gathering and making some kind of judgment on the basis of this information (Scheerens, 1983; De Groot, 1986). Involved are processes of collecting and making judgments about systems, programs, materials, procedures and processes (Harlen, 2007). A further expectation is that this "valued information" is used both for decisions on the day-to-day running of education systems or for more involving decisions on the revision and change of the system. The term *assessment* is used for student evaluation, and refers to processes in which evidence of learning is gathered in a systematic way in order to make a judgment about student learning (Harlen & Deakin Crick, 2002; OECD; 2013).

Evaluation

Traditionally, individual teachers were seen as the sole responsible agents for the quality of educational processes. Teachers were expected to (in)formally monitor pupil achievement and to adapt their instructional behavior if necessary (Faubert, 2009; Ingram, Louis & Schroeder, 2004; Slavin, 2002, Wiliam, 2011). Decentralization and accountability policies changed the demands and needs for evaluation in schools. These policies led to increasing pressure on schools to demonstrate effectiveness, facilitate school choice and to expectations about school evaluation as supporting school improvement (Devos &

[1] *"This sequence in which actions precede goals may well be a more accurate portrait of organisational functioning. The common assertion that goal consensus must occur prior to action obscures the fact that consensus is impossible unless there is something tangible around which it can occur. And this "something tangible" may well turn out to be actions already completed. Thus, it is entirely possible that goal statements are retrospective rather than prospective* (Weick, 1969, p. 8).

Verhoeven, 2003; Faubert, 2009; OECD, 2013). Systematic evaluation of teachers, school leaders, programs or the school as a whole has risen and larger and more varied use is being made of data from both internal and external evaluations (OECD, 2013). Also there is more attention to the process of using evaluative data to inform decisions at school. This process is often referred to as evidence based or data-based decision making (Ledoux, Blok, Boogaard & Krüger, 2009; Henig, 2012; Park & Datnow, 2009; Vermeulen & Van der Kleij, 2012). In data-based decision making it is recognized that decisions may be informed by various types of data (such as e.g. assessment data, process data from school self-evaluations or satisfaction data from pupil or teacher surveys), might be taken at different levels in the school (e.g. classroom, school and school board) and may lead to three types of potentially interrelated outcomes: organizational change, change in teaching and learning practices and student learning (Coburn & Turner, 2011).

Data use includes a number of processes, conditions and contexts. Central in the process is the role of interpretation, i.e. how individuals notice data, how they give meaning to it and how they construct implications for actions. The impact of the actions is subsequently evaluated by collecting new data which creates a continuous cycle of feedback and inquiry (Coburn & Turner, 2011; Mandinach & Jackson, 2012; Timperley, 2009).

Assessment

Assessment involves deciding, collecting and making judgments about evidence based on individual student progress and achievement of learning goals (Harlen, 2004; OECD, 2013). Three broad purposes of assessment in schools are to inform and support learning, to report achievement for certification, progress or transfer and to satisfy the demands of public accountability (Black, 1998)

A common distinction in the literature is that between formative and summative assessment (see e.g. Bennet, 2011; Black & Wiliam, 1998a; Harlen & James, 1997; Roos & Hamilton, 2005; Sadler, 1989, 1998; Wiliam 2011). It was Scriven (1967) who first, within the context of program evaluation, introduced the concepts summative and formative evaluation. According to Scriven (see Bennett, 2011, p. 6) 'summative evaluation provided information to judge the overall value of an educational program (as compared to some alternative), whereas the results of formative evaluation were aimed at the facilitation of program improvement'. Bloom (1969) distinguished between formative and summative evaluation in the same way, but within the context of student assessment in mastery learning. Formative evaluation then was aimed at providing feedback and correctives at each stage of the learning process, whereas summative evaluation referred to tests given at the end of an episode of teaching with the aim of grading or certifying students (Bloom, 1969; Bloom, Hasting & Madaus, 1971). In doing so Bloom mixed the purposes of assessment with the use of its results in determining whether assessment is formative or summative. Later on (see e.g. Sadler, 1989) authors used the term formative assessment instead of formative evaluation to emphasize the focus on students instead of programs.

Other scholars (Bennet, 2011; Halverson, Prichettt & Watson, 2007; Roos & Hamilton, 2005) base the distinction just on the actual use of the assessment evidence, as the same assessment instrument and evidence could be used for both summative and formative purposes. For these authors the purposes are hardly distinguishable and formative and summative assessment can coexist as primary and secondary purposes of the same assessment (Black & Wiliam, 1998a; Dunn & Mulvenon, 2009). At present there still is no clear consensus about the meaning of the term formative assessment (Sluijsmans, Joosten-Ten Brinke & Van der Vleuten, 2013; Wiliam, 2011). Formative assessment is a broad concept that covers many definitions. Brookhart (2007, p. 44) shows how the concept of formative assessment evolved in the course of time. Nowadays definitions of formative assessment could be characterized by referring to information on the learning process (Scriven, 1967), that can be used by teachers to take decisions on teaching and learning (Bloom, 1969), that actively engages students through self- and peer assessment (Sadler, 1989) and that motivates students (Black & Wiliam, 1998a, Brookhart, 2007; Crooks, 1987). In line with this evolution Black & Wiliam (2009, p. 9) defined formative assessment as follows: 'Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted and used by teachers, learners or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, then the decisions they would have taken in the absence of the evidence that was elicited'. Formative assessment according to this definition is seen as an integrated part of the teaching and learning process and not only provides teachers with information they can use to provide feedback and to improve instruction (Vermeulen & Van der Kleij, 2012) but also actively involves learners and their peers in these processes (Wiliam, 2011).

Formative assessment is primarily aimed at improving teaching and learning in the classroom and for individual pupils (Vermeulen & Van der Kleij, 2012). Formative assessment can also be applied at higher aggregation levels such as the school and above school level. In that case formative assessment provides opportunities for teachers and school leaders to learn from organizational performance data and to adjust teaching and learning processes accordingly (Halverson, Prichett & Watson, 2007; Parr & Timperley, 2008; Dunn & Mulvenon, 2009). Applied at higher aggregation levels than that of individual students or the classroom, the term formative *evaluation* rather than *assessment* is often used (see e.g. Harlen, 2007).

Feedback

Feedback is seen as a crucial component in formative evaluation and formative assessment and one of the factors that have strongest impact on student learning (see e.g. Black & Wiliam, 1998a; Crooks, 1987; Hattie & Gan, 2011; Hattie & Timperley, 2007; Shute, 2008; Stobart, 2008; Supovitz, 2012).

Shute (2008) uses the term ´formative feedback´ which she defines as information communicated to the learner that is intended to modify his or her thinking or behavior for the purpose of improving learning´ (p. 541). Feedback then is aimed at encouraging and enabling students to reduce the discrepancy between the current understanding and performance on the one hand and the desired learning goal on the other (Hattie & Timperley, 2007; Rakoczy, Harks, Klieme, Blum & Hochweber, 2013).

Hattie & Timperley (2007) developed a feedback model in which they differentiate between four types of feedback and three feedback questions (Hattie & Gan, 2011). The feedback questions refer to the goals related to the task or performance ('Where am I going?'), to the progress being made to these goal ('How am I going?') and to the activities needed to be undertaken to make better progress ('Where to next?'). The four types of feedback build on the model developed by Kluger & DeNisi (1996), and concern task feedback (clarifying and reinforcing aspects of the learning task), process feedback (feedback about the processing of the learning task), self-regulation feedback (feedback focusing on metacognitive aspects) and feedback about the self (focusing on personal attributes).

The feedback model incorporates the three broad meanings of feedback that are distinguished in the literature: the motivational meaning (feedback as praise, as a motivator to increase a general behavior), the reinforcement meaning (feedback to reward or punish a particular prior behavior) and the informational meaning (feedback to change performance in a particular direction) (Kulhavy & Wager, 1993; Nelson & Schunn, 2009). Informational feedback consists of two types of information: verification, the judgment whether something is right or wrong (often referred to as knowledge of results), and elaboration, the information needed to guide the learner into the right direction (elaborated feedback) (Goodman, Wood & Hendrickx, 2004; Kulhavy & Stock, 1989; Shute, 2008). Elaborative feedback can have many forms such as knowledge of correct response, information about the learners' thinking processes or misconceptions as well as strategic hints and cues how to proceed (Shute, 2008; Supovitz, 2012).

Other aspects of feedback distinguished in the literature refer to the function of the feedback (which could be cognitive, metacognitive and motivational) and the presentation of the feedback (Gabelica, Van den Bossche, Segers & Gijselaers, 2012; Narciss & Huth, 2004)). The latter includes among others the timing of the feedback (immediate and delayed feedback), the frequency, the way it is presented (written, verbal and/or graphically) and the way it is mediated (by the learner, peers or the teacher).

Feedback not necessarily leads to a positive reinforcement as it can be accepted, modified or rejected by the learners (Kulhavy, 1977) and interpreted in different ways and manners (Hattie, 2009). Effective and useful feedback depends on three things: motive (the leaner needs the feedback), opportunity (the learner receives the feedback in time to use it) and means (the learner is able and willing to use the feedback) (Shute, 2008; Stobart, 2008). The willingness to use feedback is related to students' motivation and recognized as an important aspect of feedback (see e.g. Mory, 2004).

The role of the teacher in providing feedback is important as well. Teachers have a choice between providing complete solutions, heavily cued hints towards the correct solution, or an adaptive "scaffolding" response, in simpler terms students receiving as much help as they would need to solve the problem on their own.

Hattie & Gan (2011) see feedback as most effective if there is a high degree of transparency about the current and desired performance by both the teacher and the student. This implies a 'need to understand feedback within the context of students' learning (with peers, with adults, alone), at varying stages of proficiency (novice, proficient, expert) and understanding (surface, deep, conceptual) with different levels of regulation (by others, with others and self) and with different levels of information and focus in the feedback information' (p. 266).

The social context in which the feedback is received is important as well. Therefore to optimize the information provided through feedback the characteristics of instruction and learners should be taken into consideration as well (Narciss & Huth, 2004; Goodman et al., 2004; Hattie & Gan, 2011).

Evaluation and Assessment as a Cyclic Process

In this review and meta-analysis assessment will be distinguished from evaluation. Assessment refers to the specific processes and tools of data collection on student progress and achievement of learning goals (see also Dunn & Mulvenon, 2009). Evaluation is a related but separate concept and concerns the processes of ascribing worth or merit to the data collected, as well as the interpretation, judgment and use of the data. In this meta-analysis evaluation is seen as a cyclic process, in which assessment could be included in a specific phase, the phase of data collection. In addition to assessment based data, other types of data (e.g. process data or student satisfaction data) could be gathered as well, during the data collection phase.

Feedback is considered a separate phase in this cycle as well and is defined as "information provided by an agent (e.g. teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" (Hattie & Timperly, 2007, p. 81). According to this definition feedback can be provided to students as educators as well.

Diverse authors proposed stage or phase models of the evaluation or assessment cycle (Birenbaum, Kimron, Shilton & Shahaf-Barzilay, 2009; Ledoux et al., 2009; Marsh, 2012; Natriello, 1987; Schuyler Ikemoto & Marsh; 2007). When comparing the phases distinguished by these authors some differences and similarities appear.

With the exception of Marsh et al. and Schuyler et al. all other authors determine '*Establishing the goals and objectives of the evaluation*' as the first phase in the cycle. Natriello add to this the setting of criteria and standards.

The second phase '*data collection'*, distinguished by all authors is the phase in which the raw data are collected. These data could be assessment based (Natriello) but could also include other type of data such as input, process and satisfaction data (other authors).

Ledoux et al. subsequently include a third phase aimed at *data administration* in which the results of the data collection need to be recorded either on paper or in an automatized computer system.

The fourth phase, *noticing and interpretation of the data,* is distinguished by all authors. This phase involves noticing the data or patterns in the data in the first place

(Coburn & Turner, 2012). Next the data must be interpreted, i.e. fitted to preexisting beliefs or cognitive frameworks and compared with the goals, criteria and standards established in the first phase (Spillane & Miele, 2007; Weick, 1995). Subsequently, by synthesizing and prioritizing the information will be transformed into actionable knowledge (Light, Wexler & Heinze, 2005) In doing so raw data are made meaningful, the gap between the intended and obtained outcome could be established and underlie decision making.

Schuyler et al., Ledoux et al. and Marsh then distinguish a fifth phase in which *decisions* on implications for actions are taken.

In a sixth phase *feedback* could be generated (Natriello; Birenbaum et al.; Marsh). This phase of the model involves the communication of the learning outcomes to the performers and stakeholders (including the learners and the teachers) as well as to provide information on how to foster subsequent teaching and learning.

Birenbaum et al., Marsh and Schuyler et al. then distinguish a seventh phase, *use,* in which the results of the evaluation are used to implementing interventions (taking actions or adjusting one's practice) by teachers and/or students to close the gaps.

In the eight phase, the *evaluation* phase, distinguished by Birenbaum et al., Marsh and Schuyler et al., it is determined in which degree the effectiveness of the interventions in closing the gaps is assessed. Judging impact requires the use of assessment information on a daily, term by-term and annual basis.

In our view a more synthetic presentation of these phases is possible. After setting the goals, data collection and administration, evaluative interpretation of the data could be mentioned, - as a third phase, and after feedback one broad category of application of the evaluative results could be distinguished, uniting categories like use, implementation, action, and decision, - as a fifth phase. After this fifth and last phase, the cycle could recommence. In summary our reformulation of the cyclic process of evaluation and feedback features the following phases: 1) setting the objectives and standards of the evaluation, 2) data collection, 3) evaluative interpretation of the data, 4) feedback and 5) use, implementation, and action.

Within the school many evaluative cycles can be going on at the same time (Vermeulen & Van der Kleij, 2012). The frequency in which these cycles are completed depends, among others, on the type of feedback and the primary audience of the data (Supovitz, 2012). Wiliam and Leahy (2006) distinguish between long, medium and short evaluative cycles. Supovitz (2012) adds a fourth category, the interim evaluation cycle. The duration of the cycles varies from one lesson (short evaluation cycle, usually based on informal classroom "checks" or assessments) to a year or more (long term evaluative cycle, based on e.g. high stakes large-scale standards based assessments or school self-evaluation results).

State of the Art: Results from Earlier Review Studies and Meta-Analyses on Evaluation and Assessment

Below we discuss the results of meta-analyses on evaluation and assessment including meta-analyses that focused explicitly on the feedback phase. An overview of studies and main effects is provided in Table 4.1.

Table 4.1
Overview of earlier meta-analyses on the effects of evaluation and assessment on student achievement

Authors	Headings of evaluation, assessment or feedback concepts in studies included	Estimated mean effect size	Number of studies included
Evaluation			
Fuchs & Fuchs (1986)	Providing formative evaluation to teachers	r = .33	21
Scheerens, Luyten, Steen & Luyten-de Thouars (2007)	Monitoring at school and class level	r = .06	43
Kyriakides, Creemers, Antoniou & Demetriou (2010)	Student assessment (school level)	r = .18	12
	Evaluation of school policy on teaching and actions taken for improving teaching practice (school level)	r = .13	6
Kingston & Nash (2011)	Formative assessment	r = .10	13
Kyriakides, Christoforou & Charalambos (2013)	Assessment (class level)	r = .34	27
Assessment			
Bangert-Drowns, Kulik, & Kulik (1991)	Frequency of classroom testing	r =. 11	35
Kim (2005)	Performance assessment	r = .17	148
Seidel & Shavelson (2007)	Evaluation of learning (class level)	r = .02	10
Hattie (2009)	Frequency or effects of testing	r = .17	569
Feedback			
Kluger & DeNisi (1996)	Feedback	r = .19	131
Hattie & Timperley (2007)	Feedback	r = .35	196

In the meta-analyses the effect sizes were expressed either as a standardized mean difference between an experimental and a control group (indicated with coefficient Cohen's *d*) or as correlations (indicated with the coefficient *r,* expressing the product moment correlation. Standardized mean differences and correlation coefficients (*r* and *d*) are convertible to one another[2]. In describing the results the effect sizes will be as correlations.

Evaluation

The impact of evaluation has been analyzed in four meta-analyses (Fuchs & Fuchs, 1986; Kyriakides et al., 2010; Kyriakides et al. 2013; Scheerens et al. 2007). With the exception of Fuchs and Fuchs these authors all focused on the impact of teaching and school factors on student achievement more broadly.

The mean effect Scheerens et al. (2007) reported for monitoring at class and school level was r = .06. Conducting moderator analyses, these authors found a significantly higher effect when mathematics or language achievement was the outcome variable (r = .23 and r = .22 respectively). Scheerens et al. applied a broad operationalization of evaluation including almost all phases of the evaluative cycle and referring both to student outcomes and process evaluation. What is more, effect sizes turned out to be weaker when monitoring was included in studies that applied multi-level modelling (r = .033) and when studies were carried out in the Netherlands (r = .02) or USA (r = .01).

Kyriakides et al. (2010) and Kyriakides et al. (2013) conducted meta-analyses on studies examining school effectiveness enhancing factors included in the dynamic model of educational effectiveness. In the 2010 meta-analysis, for evaluation and assessment, the impact of two school level variables on cognitive achievement was explored, i.e. student assessment (with an effect of r = .18) and evaluation at school level (with an effect of r = .13 reported).

The 2013 meta-analysis by Kyriakides et al. focused on the impact of teaching behaviors, i.e. the eight factors incorporated in the dynamic model of educational effectiveness. For seven of the eight factors moderate effects were found, varying from r = .34 to r = .0.46. Concerning assessment the effect was r=. 34 The effects appeared to be stronger for studies with a longitudinal design (r =. 40) and when outlier studies were removed from the sample (r = .43). In the 2013 study the operationalization of assessment included elements like using appropriate techniques to collect data on student knowledge and skills, analyzing data in order to identify student needs, reporting the results to parents and practices and evaluating their own practices (see Kyriakides, 2012).

Fuchs and Fuchs (1986) examined the impact of providing formative evaluation to and of teachers concerning the academic achievement of students. The mean effect the authors reported was r = .33. The effectiveness strongly depended on the analysis and use of the feedback by the teachers. In studies where teachers were required to apply explicit rules about the review of the student achievement data and changes in students' individualized programs to be followed, the mean effect appeared to be much higher than in studies where

[2] Converting from r to d (Borenstein, , Hedges, Higgings & Rothstein., 2009, p. 48), is as follows:

$$\mathrm{d} = \frac{2\mathrm{r}}{\sqrt{1-\mathrm{r}^2}}$$

teachers could judge themselves how to make changes in students´ individual programs (r = .41 vs r = .21 respectively).

The most commonly cited paper on the positive impact of evaluation and assessment on student achievement is the seminal review from Black and Wiliam (1998a; 1998b; 1998c). Black and Wiliam (1998a) examined 250 studies that investigated the impact of different learning strategies and approaches on student achievement. Included were studies addressing teaching and learning approaches such mastery learning or curriculum-based measurement, as well as studies that examined the effects of effective feedback, questioning, goal orientation and self- and peer assessment. The authors did not conduct a meta-analysis because of the diversity in assumptions on learning in the studies included in their paper and the lack of well-defined and widely accepted meanings of the term formative assessment. They concluded that, without doubt, formative assessment has a profound effect on learning. At the same time they argued that the theoretical basis still needs further development and attention from both researchers and practitioners (see also Bennet, 2011; Dunn & Mulvenon, 2009).

In a recently published meta-analysis on formative assessment (Kingston and Nash, 2011) the authors took a broad perspective on formative evaluation as well. A wide range of interventions at school and class level was included in the meta-analysis, including professional development with regard to formative assessment, the use of curriculum embedded assessment systems, the use of computer-based formative assessment systems and the use of feedback to students. However, the majority of effects included referred to the impact of formative assessment based on professional development. The meta-analysis yielded an average effect of r =.10, which is much smaller than the frequently cited effects of r = .20 - .33 that are often attributed to the Black and Wiliam's review.

Assessment

The association between assessment and student achievement was examined in four meta-analyses (Bangert-Drowns et al., 1991; Hattie, 2009; Kim, 2005; Seidel & Shavelson, 2007). Bangert-Drowns et al. and Hattie both summarized the effects of the frequency of testing. The average effect across eight meta-analyses that Hattie (2009) reported was r = .17, while the meta-analysis by Bangert-Drowns et al. yielded an overall effect of r = .11.

In the Bangert-Drowns et al. meta-analysis test frequency in the control group appeared to be the best predictor of effect size. The frequency of tests given to the experimental group did not influence the effect size.

Seidel and Shavelson (2007) assessed the impact of evaluation of learning. Although labeled as evaluation of learning, the evidence was limited to studies examining the impact of assessment and tests on cognitive achievement and yielded a very small effect (r = .02).
Kim (2005) finally investigated the effects of implementing a specific type of testing, which is performance assessments, on student learning. The effect Kim reported was r = .17. Performance assessment appeared to be more effective the longer it had been implemented and the more it had been integrated into instruction.

Feedback

With regard to the effects of feedback, many systematic reviews and meta-analyses are available, focusing on various aspects of feedback and conducted within a range of learning contexts (school, higher education and workplace learning). Most often, these meta-analyses were based on experimental studies (Evans, 2013) and did not lead to univocal evidence is available yet (Kluger & DeNisi, 1996; Shute, 2008). As Shute (2008, p. 157) concluded ‘the specific mechanisms relating feedback to learning are still murky, with very few general conclusions´.

One of the most influential meta-analyses on feedback is the review by Kluger & DeNisi (1996). This meta-analysis summarized the results of 131 experimental studies (470 effect sizes) on the effects of feedback interventions, many of them not classroom based. The study yielded an overall positive impact of feedback of r = .19, but also showed that in over one third of the studies feedback resulted in a negative effect and actually lowered average performance. Differential effects were found as well. Feedback on performance had highest effects when it focused on task learning and motivation (e.g. when it included information on the correct solution, or when it involved goal-setting and when the complexity of the tasks was not too high). Feedback was least effective when it focused on the self (feedback as praise).

Hattie and Timperley (2007) summarized the effects from 12 meta-analyses assessing the impact of feedback in classrooms. The average effect across the 12 meta-analyses was r = .35, which is twice the magnitude of the effect that Kluger & DeNisi reported. The effects sizes reported in the meta-analyses varied between r =.06 and r = .53, indicating that some forms of feedback are more effective than others. Just like in the Kluger & DeNisi meta-analysis the most effective types of feedback appeared to be task or process related (i.e. providing cues or reinforcement to learners), and/or related to goals. Programmed instruction, praise and punishment appeared to be the least effective forms of feedback.

From this overview on earlier reviews and meta-analyses two conclusions can be drawn: firstly, the complexity and heterogeneity of the predictor variables used in the reviews and meta-analyses stand out, secondly, there are huge differences in the effect size estimates from quantitative meta-analysis, with a noteworthy tendency for the more recent studies reporting smaller effect sizes, the Kyriakides et al. (2013) study being an exception. Finally, when comparing the effect sizes from studies which included feedback to those of studies in which the emphasis is more on just data-collection and assessment, one could conclude that feedback is a crucial link in making evaluation and assessment effective.

Method

The meta-analysis reported in this article consists of a reanalysis and extension of earlier meta-analyses published by Scheerens and Bosker (1997), Scheerens, Luyten, Steen and Luyten-de Thouars (2007) and Scheerens, Seidel, Witziers, Hendriks & Doornekamp (2005). The previous meta-analyses used studies published between 1985 and 2005. The data available from the earlier meta-analyses were combined with the data from recent studies.

Search Strategy and Selection Criteria

To select studies on evaluation at school level, evaluation at class level and assessment published in the period 2005-2010 a computer assisted search was conducted in February 2011. The following online databases were used: Web of science (www.isiknowledge.com); Scopus (www.scopus.com); and ERIC and Psycinfo (provided through Ebscohost). The databases were primarily explored using the same key terms as used in the meta-analysis by Scheerens et al. (2007): school effectiveness, educational effectiveness, teacher effectiveness, effective teaching, effective instruction, instruction, mastery learning, constructivist teaching, mathematics instruction, reading instruction, science instruction, mathematics teaching, reading teaching, science teaching. Each effectiveness keyword was crossed with each of the following output keywords: value added, attainment, achievement, learn* result*, learn* outcome*, learn* gain, student* progress and with keywords with regard to the variables of interest for this meta-analysis: feedback, evaluation, assessment, "data use", data-based*, data-driven*, reinforcement, evaluation, monitoring and test*. A total of 1105 publications matched combinations of the keywords. After removing the duplicate publications 802 publications were selected for the next step.

The titles and abstracts of each publication were screened by the reviewers. Studies were selected for closer examination if the title or abstract met the following in-and exclusion criteria:

- The study had to include an independent variable at school or class level measuring evaluation, feedback or assessment.
- The study had to include a quantitative measure of cognitive student achievement or student achievement gain of mathematics, language, science or other school subjects as the dependent variable. Examples include scores on standardized tests, achievement gain scores and grades in subject areas.
- The study had to be conducted in primary or secondary education (for students aged 6-18). Studies conducted in preschool, kindergarten or in postsecondary education were excluded from the meta-analysis.
- The study had to focus on regular students. Studies focusing on specific target groups of students in regular schools (such as students with learning, physical, emotional, or behavioral disabilities) or studies conducted in schools for special education were excluded from the meta-analysis.
- The study had to be published or presented no earlier than January 2005 and before January 2011.
- The study had to be written in English, German or Dutch.
- The study had to have estimated the relationship between a measure of evaluation or assessment and student achievement. This means that the study had to provide one or more effect sizes or had to include sufficient quantitative information to make it possible to estimate the effect size statistic.

Titles and abstracts of publications were evaluated on the selection criteria. Using the above mentioned selection criteria 255 publications published between 2005 and 2010 remained for full-text review. In addition to identify additional published studies, recent reviews and books on evaluation, feedback and assessment were examined, as well as the literature review sections from the obtained articles, chapters, research reports, conference papers and dissertations. The full text review phase resulted in 40 publications covering the period 1985-2010 to be coded or rechecked in the coding phase.

Coding Procedure

Lipsey and Wilson (2001) define two levels at which the data of the study can be coded: at study level and at the level of an effect size estimate. According to the authors a study can be defined as "a set of data collected under a single research plan from a designated sample of respondents" (Lipsey & Wilson, p. 76). A study may contain different samples, when the same research is conducted on different samples of participants (e.g. when students are sampled in different stages of schooling -primary or secondary-) or when students are sampled in different countries. An estimate is an effect size, calculated for a quantitative relationship between an independent and dependent variable.

The studies between 1985 and 2004 already had been coded. The studies selected between 2005 and 2010 were coded by the researchers applying the same standardized coding form as used by Scheerens et al. (2007). The coding form included five different sections:

- *Report and study identification.*
  This section recorded the author(s), the title and the year of the publication;
- *Characteristics of the independent (evaluation and assessment) variable(s) measured.*
  In this section the conceptualization of the evaluation and assessment variable(s) used in the study were coded. The operational definitions of the variables used in the studies were recorded too.
- *Sample characteristics.*
  The sample characteristics section recorded the study setting and participants. For study setting the country or countries included in the study were coded. With regard to participants, the stage of schooling (primary or secondary level) the sample referred to was coded as well as the grade or age level(s) of the students the sample focused on. The number of schools, classes and students included in the sample were recorded as well.
- *Study characteristics.*
  In this section the research design chosen, the type of instruments used to measure the time variable(s), the statistical techniques conducted and the model specification were coded. For research design we coded whether the study applied a quasi-experimental - or experimental design and whether or not a correlational survey design was used. With regard to the type of instruments used we coded the respondents (students, teachers, principals and/or students), whether a survey instrument or log was used and whether data were collected by means of classroom observation or video-analysis or (quasi)

experimental manipulation. The studies were further categorized according to the statistical techniques conducted to investigate the association between time and achievement. The following main categories were employed: ANOVA, Pearson correlation analysis, regression analysis, path analysis/LISREL/SEM and multi-level analysis. We also coded whether the study accounted for covariates at the student level, i.e. if the study controlled for prior achievement, ability and/or student social background.

- *Effects of evaluation and assessment (effect sizes).*
  Finally, the evaluation and assessment effects section recorded the effects sizes, either taken directly from the selected publications or calculated (see section calculation of effects sizes below). The effect sizes were coded as reflecting the types of outcome variables used (i.e. achievement test score, value-added output indicator, gain score, attainment measure, grade) as well the academic subject(s) addressed in the achievement measure. Four groups of subjects were distinguished in the coding: language, mathematics, science and other subjects. With regard to the types of outcome variables used we also coded whether the outcome variable reflected individual student level performance or whether achievement reflected a class or school mean measure.

Vote Counting Procedure

A vote counting procedure was applied to permit also inclusion of those studies that reported on the significance and direction of the association between a measure of evaluation or assessment and student achievement, but did not provide sufficient information to permit the calculation of an effect size estimate.

Vote counting comes down to counting the number of positive significant, negative significant and non-significant associations between an independent variable and a specific dependent variable of interest from a given set of studies at a specified significance level, in this case different conceptualizations of evaluation and assessment and student achievement (Bushman & Wang, 2009). We used a significance level of α=.05. When multiple effect size estimates were reported in a study, each effect was counted separately in the vote-counts. Vote counting procedures were applied for each of the three main independent variables: evaluation at school level, evaluation at class level and assessment.

The vote-counting procedure has been criticized on several grounds (Bushman, 1994; Bushman & Wang, 2009; Borenstein et al., 2009; Scheerens et al., 2005). It does not incorporate sample size into the vote. As sample sizes increase, the probability of obtaining statistically significant results increase. Next, the vote-counting procedure does not allow the researcher to determine which treatment is the best in an absolute sense as it does not provide an effect size estimate. Finally, when multiple effects are reported in a study, such a study has a larger influence on the results of the vote-count procedure than a study where only one effect is reported.

As vote counting is less powerful it should not be seen as a full blown alternative to the quantitative synthesis of effect sizes, but, rather as a complementary strategy when the information required to calculate effect sizes is missing from many studies in the sample (DeCoster, 2004; Dochy, Segers, Van den Bossche & Gijbels, 2002).

Table 4.2 provides an overview of the number of studies and effect sizes included in the vote count.

Table 4.2
Number of studies and estimates included in the vote-counting procedure

	Studies	Effect size estimates
Evaluation at school level	15	107
Evaluation at class level	25	146
Assessment	15	79

Calculation of Effect Sizes

In the majority of studies that were fully coded in our database, coefficients from regression and multilevel analysis were reported. Standardized regression coefficients were substituted directly for correlation coefficients as coefficients from multiple regression correspond to *r* equally well (for β coefficients between -.50 and .50, see Peterson and Brown, 2005). For studies that reported unstandardized coefficients, standardized coefficients were computed if the standard deviations of the explanatory variable and the achievement measure were presented in the publication. This was only possible for a minor number of studies. In these cases we applied the formulae presented in Hox (1995, p. 25) to calculate the standardized regression coefficient and standard error. For the majority of studies that reported unstandardized regression coefficients, we were not able to calculate standardized coefficients. Therefore these studies were excluded from the quantitative meta-analysis and only included in the vote counting analysis.

In some studies multiple techniques for data-analysis were applied, e.g. bivariate Pearson correlations and regression or multilevel analysis. For these studies the coefficients of the most appropriate method (regression or multilevel) were included in the meta-analysis. For studies in which bivariate or partial correlation were the only statistical techniques used or for studies for which we were not able to calculate standardized regression coefficients, the estimated Pearson correlation coefficients were included in the meta-analysis. For studies that applied regression or multilevel modeling and in which different (intermediate and final) models were presented, the coefficient(s) from the most fully identified model without interaction effects were used for the meta-analysis.

The unit of analysis for this meta-analysis was the independent sample. A study may contain different samples, when the same research is conducted on different samples of participants (e.g. when students are sampled in different grades, cohorts of students or students in different stages of schooling -primary or secondary-) or when students are sampled in different countries. In calculating the effect size in each sample, some samples reported multiple effect sizes, while other samples provide a single effect size. As inclusion of multiple effect sizes based on the same sample in one analysis violates the assumption of statistically independence (see Bennett, 2011, Cooper, Hedges & Valentine, 2009, Lipsey & Wilson, 2001), we averaged multiple effect sizes to yield a single mean effect size at sample level.

Average effect sizes were computed when:

- multiple measures or operationalizations of the same explanatory variable were included in the same analysis (e.g. evaluative feedback and informational feedback);
- multiple measures of the dependent variable were used to assess student achievement (e.g. when both a reading and writing test are used to measure language achievement or when achievement tests are used in different subjects, e.g. language and math);
- Different grade levels from the same school level were included in the analysis (e.g. both grade 4 and 6 in primary school).

Effect sizes were not averaged in the following cases:

- Analyses were performed per country in case more countries were included in a study (e.g. Swaziland and in Tanzania).
- Different school levels were included (e.g. both primary and secondary level).

Table 4.3 provides an overview of the number of studies, samples and effect sizes included in the quantitative meta-analysis. The number of estimates refers to the number of effects reported in the sample after averaging these. Due to the low number of effect size estimates for language and math separately we were not able to perform the meta-analyses also for these achievement domains separately.

Table 4.3

Number of studies and estimates included in the meta-analysis (1985-2010)

	Studies	Samples	Effect size estimates
Evaluation at school level	7	7	7
Evaluation at class level	14	15	15
Assessment	6	7	7

In order to compare the different effect size estimates used in the studies, we transformed the reported effects size estimates into Fisher's Z units using formulae presented by Lipsey and Wilson (2001).

The transformation from sample correlation r to Fishers'z is given by (see equation1)

$$z = 0.5 \text{ x } \log \left(\frac{1 + r}{1 - r}\right) \qquad (1)$$

where *z* = Fisher's z and r = correlation coefficient

The variance of *z* is given by (see equation2)

$$V_z = \frac{1}{n-3} \qquad (2)$$

In our meta-analysis we included both studies in which the outcome variable reflected individual student level performance and studies where student achievement reflected a class or school mean measure. In our study *n* referred to the number of students if the effect was related to an outcome measure based on individual student achievement and to the number of classes or schools when the effect size was related to an outcome measure based on mean class or school achievement.

Fixed and Random Effects Models

A meta-analysis can be conceptualized using a fixed-effect model or a random effects-model (Borenstein, Hedges, Higgings & Rothstein, 2010: Field & Gillet, 2010). Fixed effects models only permit inferences about the studies included in the meta-analysis, while random effects models allow generalizations to comparable studies that have been or might be conducted beyond the studies included in the meta-analysis (Field & Gillet, 2010).

Under the fixed model it is assumed that all studies in the analysis estimate the same true effect size. Under a random effects-model, the true effect size is expected to be similar but not identical across studies. The random effects model allows that there may be a distribution of true effect sizes (Borenstein et al., 2010). In a fixed model the variability between effect size estimates is due to random sampling error alone. In the random effects model the amount of variation between effect sizes is due to sampling error (the within sample variance like the fixed effects model) plus variability assumed to be randomly distributed in the population of effects (the between-sample variance) (Borenstein et al., 2009; Lipsey & Wilson, 2001). Thus, calculating the error of the mean effect size in random effects models involves estimating two error terms, whereas in fixed-effects models there is only one error term (Field & Gillet, 2010). Under the random effects model sample weights are more similar to one another than under the fixed-effect model.

As the studies in our meta-analysis vary in sample size, the effect size estimates derived from these studies differ in precision. In order to obtain the most precise effect sizes, each study is weighted by the reliability of the information. Under a fixed effects model each study is weighted by the inverse of the sampling variance (see equation 3).

$$w = \frac{1}{SE^2} \quad (3)$$

The random effects-model weights each study also by the inverse of the sampling variance plus a constant that represents the variability across population effects (see equation 4) (Borenstein et al., 2009).

$$w = \frac{1}{SE^2 + v_\phi} \quad (4)$$

where v_ϕ = the between samples variability

With the exception of the case in which the between-samples variance is zero, the variance, standard error and confidence interval for the average effect size will be wider under the random effects model.

A random-effects model is assumed most appropriate, because of the large differences in settings, designs, instrumentation, treatment and statistical techniques used in the studies. The selected studies are considered as belonging to the population of studies on the impact of evaluation and assessment on achievement.

The variability of effect sizes was investigated by applying a homogeneity test (the Q test). The *Q* statistic has an approximate chi-square distribution with *k* − 1 degrees of freedom, where k is the number of independent effect sizes. A statistically significant *Q* indicates that the variance among effect sizes is greater than can be expected by chance or sampling error alone and the null hypothesis of homogeneity of effect sizes might be rejected. This variability then can be explored by conducting further moderator analyses (Hedges & Olkin, 1985).

However, if the number of studies is small and the within-studies variance is large, the test based on the Q statistic usually has low power to detect genuine variation in population effect sizes (Field & Gillet, 2010; Hedges & Pigott, 2001). According to Hedges and Olkin (1985) the Q statistic is only accurate when the effect sizes are smaller than 1.5 and when there are more than ten observations.

Data Analysis

Data analysis was conducted in SPSS and Microsoft Excel using the procedures provided by Lipsey and Wilson (2001). The analysis procedure compromised three steps. First, the weighted mean effect sizes and confidence intervals for the random effects model were calculated. Next, a homogeneity analysis using the Q statistic was performed to examine whether there was significant variability across studies. Tables annex A1, annex A2 and annex A3 present the sample sizes and estimated effect sizes of the studies included in our meta-analysis.

Finally, we compared the findings of the random effects model with those of the fixed effects model. An advantage of the fixed effects model in comparison to the random effects model is the robustness of its estimates. By applying both fixed and random model we are able to compare the findings of the most appropriate but less robust random model to those of a less appropriate but more robust fixed model. If the findings from both approaches produce similar results, this will increase the credibility of the findings. Moreover, as the number of studies included in our meta-analysis for evaluation at school level and assessment is very small, the estimate of the between-studies variance will have poor precision (Borenstein, Hedges & Rothstein, 2007; Borenstein et al., 2010). In that case, the random effects model might still be the most appropriate, but the sample size does not allow generalizations to studies not included in the sample. In that case Borenstein et al. (2007) and Borenstein et al. (2010) suggest the fixed effects model as one the less problematic options. Under the fixed-effects model the studies included in the meta-analysis are regarded as the only studies of interest.

Results

Substantive Features of the Studies Included in Vote-Count and Meta-Analysis: Analysis of Operational Definitions of Evaluation and Assessment Used

In this study the definitions and operationalizations of evaluation and assessment used in the primary studies were categorized according to the five phases distinguished in the evaluative cycle, i.e. 1) setting the objectives and standards of the evaluation, 2) data collection, 3) evaluative interpretation of the data, 4) feedback and 5) use, implementation, and action. Each of the phases in the cycle is considered as a key element in the evaluative cycle. We were interested to see how many studies addressed the full cyclic process of evaluation and which phases were mainly covered in the studies included.

Below we present an overview on how the operationalizations of evaluation and assessment in the studies included in vote count and meta-analysis refer to the phases distinguished in the evaluative cycle. We do this separately for evaluation at school level and evaluation at class level. Studies that specifically focused on student assessment and that did not include other phases of the evaluative cycle were categorized under the variable assessment.

Evaluation at School Level

Fifteen studies included in vote count and meta-analysis examined the impact of evaluation and assessment at school level. The first phase*, setting the objectives and standards of the evaluation* was addressed in two studies (Brandsma, 1993; Hofman, 1993). In these studies the goals referred to goals stated in the school work plan as guiding principle for the evaluation. The phase of *data collection* was addressed in seven studies studies (Brandsma, 1993; Creemers & Kyriakides, 2010; Hofman, 1993; Kyriakides, 2005; Reezigt, Guldemond & Creemers, 1999; Senkbeil, 2006; Yelton, Miller & Ruscoe, 1994) and referred to data collection based on student achievement as well as data collection with regard to school processes such as e.g. the evaluation of the curriculum. Operationalizations concerned both the evaluation methods applied (such as the -standard- use of achievement tests, e.g. criterion referenced tests, curriculum-dependent tests) as well as the procedures for data collection (e.g. the yearly evaluation of educational activities, the standardization of testing procedures) and the registration of the data (i.e. the systematic registration and documentation of pupil progress). The third phase, *evaluative interpretation of the data*, was addressed in only one study (Bosker & Hofman, 1987) and concerned the way in which teachers conduct error analyses. Three studies examined the *feedback* phase (Hammond & Yeshanew, 2007; Van der Grift, Houtveen & Vermeulen, 1997; Vermeulen, 1987). In one study it was examined whether student results are used on a regular basis to inform pupils who lag behind. The second study examined the feedback of test results from the department leader to the teacher. The third study addressed the issue of paid versus unpaid school performance feedback from a national data set. *Use, implementation, and action*, the last phase, finally was addressed in seven studies (Brandsma, 1993; Creemers & Kyriakides, 2010; Hofman, 1993; Kyriakides, 2005; Schildkamp, Visscher & Luyten, 2009; Van der Grift et al., 1997; Vermeulen, 1987). Facets like: taking actions based on error analyses, improvement of teaching and/or educational program based on assessment data or data from evaluations of educational activities, using achievement data for diagnostic and remedial teaching, and checking goal attainment, are operationalized in six of the seven studies. Schildkamp et al. (2009) finally, specifically focused on the use of the results of school self-evaluation. Following Weiss (1998) in this study a distinction was made between conceptual use and instrumental use, i.e. whether the feedback from the school self-evaluation results influenced thinking of school staff (such as provided school staff with new insights or highlighted problems) or whether school staff used the results in a direct way (e.g. by taking measures to improve the quality of education).

Evaluation at Class Level

The impact of evaluation, assessment and feedback at classroom level was analyzed in 25 studies included in the vote count and meta-analysis. The first phase, *setting the objectives and standards of the evaluation* was examined in four studies (Clausen, 2001; Gruehn, 1995; Klieme & Rakoczy, 2003; Kunter, 2004; Levacic, Steele, Smees & Malmberg; 2003). Goals in these studies all referred to the ‘diagnostic competences’ of the teachers and the way they set goals and reference norms for individual students instead of using group norms. *Data collection and registration* was addressed in ten studies (Bourke, 1986; Brandsma, 1993; Clausen, 2001; Driessen & Sleegers, 2000; Gruehn, 1995; Reezigt, 1993; Reezigt et al., 1999; Senkbeil, 2006; Van der Grift et al., 1997; Van der Werf, Creemers & Guldemond, 2001). At class level the operationalization of data collection and registration in almost all studies referred to assessment (monitoring of student) work and included both short (i.e.

monitoring classroom assignments and homework) and medium and interim cycles of assessment practices (referring to testing and diagnostic testing), as well as the registration of pupil progress. Evaluation and the procedures for evaluation at class level were the independent variables in one study, although not further operationalized by the authors. *Evaluative interpretation of the data* again appeared to be a neglected phase. At class level this phase was addressed in none of the studies included in our review and meta-analysis. This in contrast to the next phase in the cycle, *feedback*, which was addressed in two third of the studies included in our review (Brandsma, 1993; Carpenter, Pashler & Cepeda, 2009; De Fraine, Van Damme, Van Landeghem, Opdenakker & Onghena, 2003; Hill & Rowe, 1998; Hofman, Hofman & Guldemond, 1999; Klieme & Rakoczy, 2003; Kyriakides, 2005; Kyriakides & Creemers, 2008; Levacic et al.; 2003; Lockheed & Longford, 1991; Rakoczy, Klieme, Bürgermeister & Harks, 2008; Reezigt, 1993; Reezigt et al., 1999; Senkbeil, 2006; She & Fisher, 2002; Van der Grift et al., 1997). Various aspects of feedback were covered in the primary studies, such as the primary users of the feedback, the types of feedback, as well as the timing of feedback (immediately versus delayed). In most studies students appeared to be the primary target group of feedback while in a few studies the feedback was (also) directed at the teachers. Encouragement and praise, i.e. rewarding pupils working hard or making good progress, verification (knowledge of correct results, knowledge of incorrect results), and elaboration (information needed to improve the achievement) are the types of feedback addressed in the studies. The last phase, *use, implementation, and action*, was examined in eight studies (Driessen & Sleegers, 2000; Hofman, Hofman & Guldemond, 1999; Levacic et al.; 2003; Reezigt et al., 1999; Rymenans, Geusdens, Coucke, Van den Bergh & Daems, 1996; Senkbeil, 2006; Van der Grift et al., 1997; Van der Werf et al., 2001). In these studies use often implied the implementation of diagnostic practices for weaker students, such as motivating underachieving students, making available remedial material, and preventing or combatting learning problems. In some (other) studies, however, use also referred to adjustment or improvement of learning or teaching goals or practices based on evaluative information.

Two studies at classroom level finally included the whole evaluation cycle (Ysseldyke & Bolt, 2007; Ysseldyke & Tardrew, 2007). These studies examined the impact of the implementation of a computer-based formative assessment system, Accelerated Math. This system keeps track of individual students' daily activities, provides immediate feedback to students and teachers, alerts teachers when students have difficulties with certain math assignments, and monitors student achievement and provide teachers with the information they need to differentiate and adjust instruction (Ysseldyke & Tardrew, 2007, p. 5).

Assessment

Fifteen studies solely focused on student assessment. The most often used operationalization of assessment refers to the frequency of testing (Bosker, Kremers & Lugthart, 1990; Driessen & Sleegers, 2000; Kyriakides & Creemers, 2008; Reezigt, 1993; Reezigt et al.; 1999; Rymenans et al., 1996; Van der Werf et al., 2001). Next to this, studies assessed the impact of the type of assessments (such as e.g. oral tests, method-dependent assessments, method-independent assessments, diagnostic tests, tests as part of a student monitoring system (Bourke, 1986; Bosker et al., 1990; Carpenter et al., 2009; Kyriakides & Creemers, 2008; Meijnen, Lagerweij & De Jong, 1993; Olina & Sullivan, 2002). The average time (per week) teachers use for assessment (Lockheed & Komenan, 1989; Schaub & Baker, 1991) or the time teachers spent on scrutinizing of tests (Pugh & Telhaj, 2003) is a third category of operationalizations of assessment that are used in the studies included in the review. Willms and Somers (2001) finally used a very basic operationalization, i.e. whether pupils are tested or not.

Based on the operationalizations of evaluation and assessment presented above it can be concluded that the full cyclic process of evaluation and assessment starting with goal setting and ultimately leading to decisions and actions to enhance teaching and learning is addressed in just a few studies. Instead for studies targeted at evaluation at school level the phases most frequently addressed refer the data collection phase and the phase of use, implementation and action. Studies examining the impact of evaluation at class level focused on these phases as well, but also examined the impact of aspects of feedback. Both at class and school level, there appeared to be less attention for the phase of goal setting, while the phase of interpretation of the data did receive attention in just one study.

Results of the Vote Counting

The results of the vote count analyses provide a rough overall picture on the question to what extent evaluation and assessment are positively related with student achievement. Table 4.4 shows the results of the vote count for evaluation at school level, evaluation at class level and assessment.

Table 4.4
Results of vote counts examining the number and percentage of negative, non-significant and positive effects of evaluation at school level, evaluation at class level and assessment on academic achievement

Conceptualization	Negative effects N (%)	Non-significant effects N (%)	Positive effects N (%)
Evaluation at school level	1 (1%)	57 (53%)	49 (46%)
Evaluation at class level	8 (6%)	100 (72%	38 (22%)
Assessment	3 (4%)	62 (78%)	14 (18%)
Total	12 (4%)	219 (68%)	91 (28%)

The vote count shows a mixed picture. On average, two third of the associations between evaluation and assessment and achievement appeared to be non-significant. Less than one third of the estimates showed positive and significant effects. The proportion of positive effects found is largest in studies that examined the impact of evaluation on school level. The results show that for evaluation at school level, the number of positive and non-significant effects do not differ substantially from each other.

The balance of negative and positive effects, when totaling the three forms of evaluation and assessment (4% versus 28%) might be seen as weak evidence for the predominance of positive effects.

Meta-Analysis

Coefficients Required for the Meta-Analysis
Tables 4.4, 4.4.5 and 6 show the, unweighted correlation coefficients, effect sizes transformed into Fisher Z coefficients and standard errors that were calculated for all studies included in the meta-analyses.

For evaluation at school level the effect sizes ranged from -.065 to .273 (see table 5 and annex A1). In two of the seven studies (Bedford, 1988; Bosker & Hofman, 1987) a negative effect of evaluation at school level was reported. Relatively large positive effects were found in three studies (Hofman, 1993; Vermeulen, 1987; Yelton et al., 1994). This might be partly due to the lack of control for student background characteristics in these studies as the statistical technique applied was Pearson correlation. Hofman et al. and Yelton et al. applied other techniques (regression and path analysis respectively) as well but the available data did not allow the calculation of standardized effects and could therefore not be included in our meta-analysis.

Table 4.5
Meta-analysis coefficients (Evaluation at school level)

Evaluation at school level	Sample	Correlation coefficient	Fisher's Z (F_z)	SE_z	95% confidence interval for Fisher Z	
Authors					Lower bound	Upper bound
Bedford (1988)		-.030	-.030	.122	-0.269	0.209
Bosker & Hofman (1987)		-.065	-.065	.053	-0.169	0.038
Creemers & Kyriakides (2010)		.062	.062	.020	0.023	0.101
Hofman (1993)		.211	.214	.018	0.183	0.245
Kyriakides (2005)		.066	.066	.024	0.019	0.113
Vermeulen (1987)		.250	.255	.267	-0,268	0.778
Yelton et al. (1994)		.267	.273	.147	-0.015	0.561
Summary effect		.117	.118			

For evaluation at class level the effect sizes ranged from -.125 to .236 (see Table 4.6 and Table A2). The largest negative effect reported resulted from the study by She & Fisher (2002). These authors examined the effects of teachers' communication behavior on cognitive and attitudinal outcomes of students in Taiwan.

Table 4.6
Meta-analysis coefficients (Evaluation at class level)

Evaluation at class level	Sample	Correlation coefficient	Fisher's Z (F_z)	SE_z	95% confidence interval for Fisher Z	
Authors					Lower bound	Upper bound
Bourke (1986)		.020	.020	.129	-0.233	0.273
Carpenter et al. (2009)		.232	.236	.130	-0.019	0.491
Clausen (2001)		-.020	-.020	.141	-0.296	0.256
Gruehn (1995)		.146	.147	.086	-0.022	0.316
Klieme & Rakoczy (2003)		.070	.070	.050	-0.028	0.168
Kunter (2004)		-.045	-.045	.023	-0.090	0.000
Kyriakides (2005)		.094	.094	.024	0.048	0.141
Kyriakides & Creemers (2008)		.063	.063	.020	0.024	0.102
Lockheed & Longford (1991)		.095	.095	.022	0.052	0.138
Rakoczy et al. (2008)		.007	.007	.065	-0.120	0.130
Reezigt (1993)		.023	.023	.015	-0.006	0.005
She & Fisher (2002)		-.124	-.125	.030	-0.184	0.066
Ysseldyke & Bolt (2007)		.190	.192	.028	0.137	0.247
Ysseldyke & Tardrew (2007)	P	.188	.191	.055	0.083	0.299
	S	.188	.191	.152	-0.107	0.489
Summary effect		.087	.087			

P= primary education, S = secondary education

The large negative effect of encouragement and praise on achievement was not statistically significant however, while the effects reported for attitudinal outcomes were positive and statistically significant. The studies that reported relatively large positive effects all had a (quasi-) experimental research design and addressed the timing of feedback (Carpenter et al., 2009) respectively the impact of a progress monitoring and instructional management system (Ysseldyke & Bolt, 2007; Ysseldyke & Tardrew, 2007).

Concerning Assessment the relatively large negative effect reported derived from the study by Bourke (1986) (see Table 4.7 and Table A3). In this study with a correlational research design Pearson correlation was the only statistical method used. Carpenter et al. (2009) on the other hand reported a relatively large positive effect of testing. The study by Carpenter at el. employed an experimental research design and examined the impact of feedback and re-study over just re-study on long-term retention of course knowledge.

Table 4.7
Meta-analysis coefficients (Assessment)

Assessment	Sample	Correlation coefficient	Fisher Z (F_z)	SE_z	95% confidence interval for Fisher Z	
Authors					Lower bound	Upper bound
Bourke (1986)		-.170	-.172	.129	-.425	.081
Carpenter et al. (2009)		.203	.206	.128	-.045	.457
Kyriakides & Creemers (2008)		.015	.015	.020	-.024	.054
Lockheed & Komenan (1989)	Nigeria	.026	.026	.038	-.048	.100
	Swaziland	.079	.079	.041	-.159	.001
Pugh & Telhaj (2003)		.028	.028	.014	.001	.055
Reezigt (1993)		.001	.001	.015	-.028	.030
Summary effect		.008	.008			

Computation of Average Effect Sizes

Application of a random effects model resulted in weighted mean effect sizes of F_z =.073 for evaluation at school level, F_z =.073 for evaluation at class level and F_z =.005 for assessment (see also Table 4.8). The results of the *z* tests showed that evaluation at school level and evaluation at class level deviated significantly from zero: for evaluation at school level *z* = 3.78 (p<.05.), for evaluation at class level *z* = 2.92 (p<.05). For assessment the results of the *z* test showed that the effect size did not differ significantly from zero (assessment *z* = 0.384, p>.05). Evaluation at school level and evaluation at class level thus seemed to have small but statistically significant effects on student achievement, while the effect of assessment was almost zero and non-significant.

Table 4.8
Overall effect sizes (fixed effects model and random effects model)

				95% confidence interval		Test of heterogeneity in effect sizes		
	k	ES	SE	Lower bound	Upper bound	Q_T value	df (Q)	p-value
Evaluation at school level								
1. Fixed	7	.070***	.015	.041	.099	8.99	7	>.05
2. Random	7	.073**	.019	.035	.111			
Evaluation at class level								
1. Fixed	15	.058***	.009	.040	.076	72.63	14	< .05
2. Random	15	.073***	.025	.024	.122			
Assessment								
1. Fixed	7	.012	.009	-.006	.030	11.41	6	>.05
2. Random	7	.005	.013	-.020	.030			

k = number of samples included in the meta-analysis, ES = weighted effect size, SE = standard error
* = significant at .05, **= significant at .01, *** = significant at .001

We then conducted the Q test of the homogeneity of the effect sizes. If the Q statistic is significant it can be assumed that the individual effect sizes differ from the population mean by more than sampling error alone. Moderator analyses then are required to determine the source of variation in addition to the sampling error (Lipsey & Wilson, 2001).

In our study the Q test proved to be statistically significant for evaluation at class level (Q_T = 72.63, df = 14, p < .05) (see Table 4.8). This means that for evaluation at class level the homogeneity analysis shows that the observed variation in the distribution of effect sizes is more heterogeneous than might be expected from sampling variance only, which indicates that conducting further moderator analyses might be appropriate. However, despite this indication it was ultimately decided not to conduct a moderator analysis given the small number of effect sizes included in the sample (n = 15).

For evaluation at school level (Q_T = 8.99, df = 6, p >.05) and for assessment (Q_T = 11.41, df = 6, p >.05) the Q statistic appeared to be statistically non-significant. This might be due to the small samples with effect sizes (n < 10) for these two variables included in the meta-analysis (n = 7 for evaluation at school level and n = 7 for assessment. Conducting further moderator analyses thus is not appropriate for these two variables as was also the case for evaluation at class level (see Borenstein et al. 2007; Hedges & Olkin, 1985).

In order to verify the robustness of our analyses the results obtained by testing random effect models, were compared to the outcomes to the testing of fixed effect models.

Table 4.8 shows the results of the comparison between the two approaches. The findings indicate that the weighted mean effects for evaluation at school level and assessment are almost the same under both the random effects model and fixed effects model. For evaluation at class level the mean effect size appeared to be somewhat lower under the fixed effects model.

Discussion

The aim of this review study and meta-analysis was to summarize and update the research on the impact of evaluation and assessment as an effectiveness enhancing school and classroom level variable on student achievement. The meta-analysis included 7 samples on evaluation at school level, 15 samples on evaluation at class level and 7 samples examining the impact of assessment. A vote count procedure was applied as well to permit the inclusion of studies that did not provide sufficient information to calculate an effect size.

Following basic elements of evaluation and assessment distinguished in the literature an "evaluative cycle" was used to cadre the conceptualizations and operationalizations of the predictor variables used in the primary studies. Five phases were distinguished: 1) setting the objectives and standards of the evaluation, 2) data collection, 3) evaluative interpretation of the data, 4) feedback and 5) use, implementation, and action. A thorough and complete application of the consecutive phases of the evaluative cycle was rarely included in the studies that were used for our review and meta-analysis. Instead, the operationalizations used in the primary studies referred to rather fragmented and often superficial measurements of activities related to one or more phases of the cycle, mainly asking for teacher and school leader self-perceptions about how frequent various evaluation, assessment, and feedback practices were applied.

Data collection and use, implementation and action were the phases most commonly addressed (in studies conducted at both class and school level), as well as the phase of feedback for evaluation (class level only). Goal setting and evaluative interpretation of data on the other hand were the phases (hardly) covered. As goal setting and achievement orientation usually are considered to be a separate effectiveness enhancing factors in school and teaching effectiveness research, this might be an explanation for the finding that this phase was hardly addressed. A similar kind of reasoning does not apply to the neglect of the phase of noticing and interpretation of the data. Although very central in the evaluative cycle, we found hardly any empirical research on the impact of teachers' and school leaders' practices and abilities to interpret student work or other data. The latter finding is touched upon also by other authors (see e.g. Bennett, 2011) who suggests that interpreting or making inferences is only just beginning to become integrated into definitions of formative assessment. As understanding students' work correctly is a crucial precondition for providing relevant feedback or adjusting teaching and learning, more attention for this phase, both in research and practice, seems to be recommendable.

The impact of evaluation at school level, evaluation at class level and assessment on student achievement was examined by means of vote counts and meta-analysis. Across the three variables, the vote counts indicated a weak general predominance of positive effects compared to negative effects (28% versus 4%), with a substantial higher percentage of positive effects for evaluation at school level. The meta-analyses yielded weak and significant positive effects for evaluation at school level and evaluation at class level (F_z= .070 and F_z = .073 respectively[3]), while for assessment the effect found was almost zero and non-significant (F_z = 0.01). The effects found in this study confirm the findings of the previous meta-analysis on which this study builds (Scheerens et al., 2007). The effects are somewhat larger than the effects found by Seidel and Shavelson, but smaller than those reported in the 2010 meta-analysis by Kyriakides et al. (r = .13 and .18) and much smaller than the effects found in the 2013 study by Kyriakides et al. (r =.34). The differences might be partly due to the studies included in the meta-analyses. Comparing the studies included in our meta-analysis with those included in the work by Kyriakides et al. (2013) shows that less than one third of the studies included in our vote count (and an even smaller number of studies in our meta-analysis) were also incorporated in the study by Kyriakides et al. The small overlap might be due to differences in the selection of the studies, (depending partly on the application of inclusion criteria), and the calculation of the effect sizes. Deeper analysis of the differences in results among meta-analyses is hindered by the fact that many of the earlier publications (Kyriakides et al., 2013, among them) neither provide a summary of the individual studies incorporated, nor a listing of the effect sizes for each study.

A similarity between our meta-analysis and the study by Kyriakides at al. (2013) is the mixed sample of research designs included (although the number of (quasi-) experimental and longitudinal studies is limited in both meta-analyses). This might not be the case for some of the meta-analyses that focused on feedback specifically. As most of the studies included in our meta-analysis were employed in actual schools and classrooms and had a correlational research design, studies included in some of the previous meta-analyses on feedback (among others Kluger & DeNisi, 1996) were mainly conducted in laboratory settings. An advantage of studies conducted in a natural setting is its stronger ecological validity, which contributes to the generalizability of findings. A disadvantage, however, might be a reduced internal validity (and lower effects) as it is not possible to have rigorous control for confounding variables.

A drawback of our meta-analysis is the small number of primary studies incorporated, especially for evaluation at school level and assessment. For these variables, the average effects reported should be interpreted with caution. Many studies examining the impact of formative assessment or data based decision making lacked sufficient information needed to compute standardized effect sizes. Therefore only half of the studies on assessment that reported coefficients form regression or multilevel analysis could be incorporated in the meta-analysis. Authors of primary studies therefore should be requested to provide standardized effects or sufficient quantitative data so that a standardized effect could be calculated.

Effectively engaging in the evaluative cycle of inquiry rests on a number of assumptions, i.e. that the data collected are accurate, that teachers and school leaders have the necessary skills to analyze and interpret the data effectively, that educators are able to provide relevant feedback and that teachers and school leaders are able to make appropriate

[3] For small values (r <.25) r equals F_z

adaptations to teaching and learning (see e.g. Schneider & Gowan, 2013). The quality of the evaluative cycle and its impact on teaching and learning rests in part on the attitudes, knowledge and skills that teachers and school leaders have in evaluation and assessment and the strategies they use. Creating effective evaluation and assessment cycles at all educational levels requires capacity building and professional learning at both teacher, school and above school level. At present there is some evidence on the impact of professional learning and other support activities on teachers' and school leaders' skills and knowledge to analyze and interpret performance data (see e.g. the studies by Christoforidou, Kyriakides & Panayiotis, 2014; Staman, Visscher & Luyten, 2014; Vanhoof, Verhaeghe, Verhaeghe, Valcke & Van Petegem, 2011). There is still limited empirical evidence about how to professionalize and support teachers in taking effective interventions in the phases of feedback and action, use and implementation. Further research therefore is recommended to understand what types of professional development and support will enhance effective evaluation and assessment practices, in particular also those interventions that are aimed at improving teachers' and school leader abilities to provide feedback and adapt instruction based on student assessment data.

A further need is also to evaluate the conditions that influence the implementation and effects of evaluation and assessment practices. Research has identified a large number of factors including intervention characteristics (data infrastructure and initiatives, tools, quality and type of data), school organization and political context characteristics (e.g. time, leadership, power relations, evaluation culture, vision, norms and goals, training and support, ownership and autonomy), user relationships and characteristics (trusts, beliefs, knowledge and skills) However, further research in this area is necessary as the amount and quality of the evidence varies and there is still limited guidance about how these factors interact: ".. *because so little of this research employs strong theoretical frameworks, we know little about how these myriad contextual factors interact with each other. Stronger theories of context could help to build knowledge across studies, interpret or explain findings, highlight relationships that persist over time and suggest causal mechanisms* (Turner & Coburn, 2012, p. 5).

References

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213-223. doi:10.3102 /00346543061002213

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*, 5-25. doi:10.1080/0969594X.2010.513678

Birenbaum, M., Kimron, H., Shilton, H., & Shahaf-Barzilay, R. (2009). Cycles of inquiry: Formative assessment in service of learning in classrooms and in school-based professional communities. *Studies in Educational Evaluation*, *35,* 130-149. doi:10.1016/j.stueduc .2010.01.001

Black, P. (1998). *Testing: Friend or foe? Theory and practice of assessment and testing*. London: Falmer.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*, 7-74. doi:10.1080/0969595980050102

Black, P., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. London, England: King's College. Retrieved from: *weaeducation.typepad.co.uk /files/blackbox-1.pdf*

Black, P., & Wiliam, D. (1998c). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. doi:10.1007/s11092-008-9068-5

Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. In R. W. Tyler (Ed.), *Educational evaluation: New roles, new means. The 63rd yearbook of the National Society for the Study of Education, part 2 (Volume 69)* (pp. 26-50). Chicago, IL: University of Chicago Press.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning.* New York, NY: McGraw-Hill.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Borenstein, B. Hedges, L. V. Higgings, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*, 97-111. doi:10.1002/jrsm.12

Borenstein, M. Hedges, L., & Rothstein, H. (2007). *Meta-analysis: fixed effect vs. random effects*. Retrieved from www.Meta-Analysis.com

Brookhart, S. M. (2007). Expanding views about formative classroom assessment: A review of the literature. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 43-62). New York, NY: Teachers College Press.

Bushman, B. J. (1994). Vote-Counting procedures in meta-analysis. In H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 193-213). New York: Russell Sage Foundation.

Bushman, B. J., & Wang, M. C. (2009). Vote-Counting procedures in meta-analysis. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.) (pp. 207-220). New York: Russell Sage Foundation.

Christoforidou, M. Kyriakides, L., & Panayiotis, A. (2014). Searching for stages of teacher's skills in assessment. *Studies in Educational Evaluation, 40*, 1-11. doi:10.1016/j.stueduc .2013.11.006

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research and Perspectives, 9*, 173-206. doi:10.1080 /15366367.2011.626729

Coburn, C. E., & Turner, E. O. (2012). The practice of data use: An introduction. *American Journal of Education, 118*, 99-111. doi:10.1086/663272

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.) (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

Crooks, T. J. (1987). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*, 438-481. doi:10.3102/00346543058004438

De Groot, A. D. (1986). Is de kwaliteit van het onderwijs te beoordelen? In A.D. de Groot, *Begrip van evalueren*. Den Haag: VUGA.

DeCoster, J. (2004). *Meta-analysis notes.* Retrieved August 26, 2013 from http://www.stat-help.com/notes.html

Devos, G., & Verhoeven, J. C. (2003). School self-evaluation: conditions and caveats. The case of secondary schools. *Educational Management and Administration*, *31*, 404-420. doi:10.1177/0263211X030314005

Dochy, F., Segers, M., Van den Bossche, P., & Gijbels, D. (2002). Effects of problem based learning: A meta-analysis. *Learning and Instruction, 13*, 533–568. doi:10.1016/S0959-4752(02)00025-7

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence on the impact of formative assessment in education. *Practical Assessment, Research & Evaluation*, *14*(7), 1-11. Retrieved from http://pareonline.net/pdf/v14n7.pdf

Edmonds, R. R. (1979). Effective schools for the urban poor. *Educational Leadership*, *37*, 15-27.

Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research, 83*, 70-120. doiI:10.3102/0034654312474350

Faubert, V. (2009). School evaluation: Current practices in OECD countries and a literature review. OECD Education Working Paper No. 42. Paris: OECD.

Field, P. F., & Gillet, R. (2010). Expert tutorial. How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *63*, 665-694. doi:10.1348 /000711010X502733

Fuchs, L. S., & Fuchs, D. (1986). Effects of systemic formative evaluation: A meta-analysis. *Exceptional Children, 53*(3), 199-208.

Gabelica, C., Van den Bossche, P., Segers, M., & Gijselaers, W. (2012). Feedback, a powerful lever in teams: A review. *Educational Research Review*, *7*, 123-144. doi:10.1016/j.edurev.2011.11.003

Goodman, J. S., Wood, R. E., & Hendrickx, M. (2004). Feedback specificity, exploration, and learning. *Journal of Applied Psychology*, *89,* 248-262. doi:10.1037/0021-9010.89.2.248

Halverson, R., Prichett, R. B., & Watson, J. G. (2007). *Formative feedback systems and the new instructional leadership*. Madison, WI: University of Wisconsin.

Harlen W. (2004). A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes. *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Harlen, W. (2007). *The quality of learning: Assessment alternatives for primary education*. *Interim Reports*. Cambridge, UK: University of Cambridge. Retrieved from http://gtcni.openrepository.com/gtcni/bitstream/2428/29272/2/Primary_Review_Harlen_3-4_briefing_Quality_of_learning_-_Assessment_alternatives_071102.pdf

Harlen W., & Deakin Crick, R. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning (EPPI–Centre Review). *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Harlen, W., & James, M. (1997). Assessment and Learning: differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, *4*, 365-379. doi:10.1080/0969594970040304

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

Hattie, J., & Gan, M. (2011). Instruction based on feedback. In P. Alexander & R. E. Mayer (Eds.), *Handbook of research on learning and instruction* (pp. 249-271). New York, NY: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112. doi:10.3102/003465430298487

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6,* 203-217. doi:10.1037//1082-989X.6.3.203

Henig, J. R, (2012). The politics of data use. *Teachers College Record, 114*(11), 1-32.

Hox, J. J. (1995). *Applied multilevel analysis* (2nd ed.). Amsterdam: TT-publikaties.

Ingram, D., Seashore-Louis, K., & Schroeder, R.G. (2004). Accountability policies and teacher decision making: Barriers to the use of data to improve practice. *Teachers College Record, 106*(6), 1258-1287. Retrieved from http://www.tcrecord.org

Kim, S-E. (2005). *Effects of implementing performance assessments on student learning: meta-analysis using HLM*. (Unpublished Doctoral dissertation). University Park, PA: The Pennsylvania State University.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*, 28-37. doi:10.1111/j.1745-3992.2011 .00220.x

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254-284. doi:10.1037/0033-2909.119.2.254

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47*, 211-232. doi:10.3102/00346543047002211

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1,* 279-308. doi:10.1007/BF01320096

Kulhavy, R. W., & Wager, W. (1993). Feedback in programmed instruction: Historical context and implications for practice. In J. V. Dempsey & G. C. Sales (Eds.), *Interactive instruction and feedback* (pp. 3-20). Englewood Cliffs NJ: Educational Technology.

Kyriakides, L. (2102). Advances in school effectiveness theory. In Ch. Chapman, P. Armstrong, A. Harris, D. Muijs, D. Reynolds & P. Sammons (Eds.), *School effectiveness and improvement research, policy and practice: Challenging the orthodoxy?* (pp. 44-57). London and New York: Routledge.

Kyriakides, L., Creemers, B., Antoniou, P., & Demetriou, D. (2010). A synthesis of studies searching for school factors: implications for theory and research. *British Educational Research Journal*, *36*, 807-830. doi:10.1080/01411920903165603

Kyriakides, L., Christoforou, C., & Charalambous, C. L. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education*, *36*, 143-152. doi:10.1016/j.tate.2013.07.010

Ledoux, G., Blok, H., Boogaard, M., & Krüger, M. (2009). *Opbrengstgericht werken over de waarde van meetgestuurd onderwijs*. Amsterdam: SCO-Kohnstamm Instituut.

Light, D. Wexler, D. H., & Heinze, J. (2005). Keeping teachers in the center: A framework of data-driven decision-making. Retrieved from http://www.edc.org/

Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Mandinach, E. B., & Jackson, S. S. (2012). *Transforming teaching and learning through data-driven decision making*. Thousand Oaks: Corwin.

Marsh, J. E. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record*, *114*(11), 1-48.

Mory, E. H. (2004). Feedback research revisited. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 745-783). Mahwah: Erlbaum.

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multimedia learning. In H. M. Niegemann, D. Leutner & R. Brunken (Eds.), *Instructional design for multimedia learning* (pp. 181-195). Munster, NY: Waxmann.

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist, 22*, 155-175. doi:10.1207/s15326985ep2202_4

Nelson, N. M., & Schunn, C. D. (2009). The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science, 37*, 375-401. doi:10.1007 /s11251-008-9053-x

OECD (2013). *Synergies for better learning: An international perspective on evaluation and assessment. OECD Reviews of evaluation and assessment in education*. Paris: OECD Publishing. doi:10.1787/9789264190658-en

Park, V., & Datnow, A. (2009). Co-constructing distributed leadership: District and school connections in data-driven decision-making. *School Leadership and Management, 29*, 477-494. doi:10.1080/13632430903162541

Parr, J. M., & Timperley, H. S. (2008). Teachers, schools and using evidence: considerations of preparedness. *Assessment in Education, Principles, Policy & Practice*, *15*, 57-71. doi:10.1080/09695940701876151

Peterson, R. A., & Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology, 90*, 175-181. doi:10.1037/0021-9010.90.1.175

Rakoczy, K., Harks, B. Klieme, E., Blum, W., & Hochweber, J. (2013). Written feedback in mathematics: Mediated by students' perception, moderated by goal orientation. *Learning and Instruction, 27*, 63-73. doi:10.1016/j.learninstruc.2013.03.002

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, Ch., & Stringfield, S. (2014). Educational effectiveness research (EER): A state-of-the-art review. *School Effectiveness and School Improvement, 25*, 197-230. doi:10.1080/09243453 .2014.885450

Roos, B., & Hamilton, D. (2005). Formative assessment: a cybernetic viewpoint. *Assessment in Education: Principles, Policy & Practice, 12*, 7-20. doi:10.1080 /0969594042000333887

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-144. doi:10.1007/BF00117714

Sadler, D. R. (1998). Formative assessment: revisiting the territory. *Assessment in Education*, *5*, 77-84. doi:10.1080/0969595980050104

Scheerens, J. (1983). *Evaluatie-onderzoek en beleid. Methodologische en organisatorische aspecten*. (Doctoral dissertation) Harlingen: Flevodruk b.v.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Elsevier Science Ltd.

Scheerens, J., Glas, C., & Thomas, S. (2003). *Educational Evaluation, Assessment and Monitoring*. Lisse: Swets & Zeitlinger.

Scheerens, J., Luyten, H., Steen, R., & Luyten-de Thouars, Y. (2007). *Review and meta-analyses of school and teaching effectiveness*. Enschede: Department of Educational Organisation and Management, University of Twente.

Scheerens, J., Seidel, T., Witziers, B., Hendriks, M., & Doornekamp G. (2005). *Positioning and validating the supervision framework.* Enschede: University of Twente, Department of Educational Organization and Management.

Schuyler Ikemoto, G., & Marsh, J. A. (2007). Cutting through the "data-driven" mantra: Different conceptions of data-driven decision making. *Yearbook of the national society for the study of education, 106*(1), 105-131.

Scriven, M. (1967). *The methodology of evaluation.* Washington, DC: American Educational Research Association.

Seidel, T., & Shavelson, R. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. R*eview of Educational Research*, 77, 454-499. doi:10.3102/0034654307310317

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153-189. doi:10.3102/0034654307313795

Slavin, R. E. (2002). Evidence-based education policies: transforming educational practice and research. *Educational Researcher, 21*, 15-21. doi:10.3102/0013189X031007015

Schneider, M. C., & Gowan, P. (2013). Investigating teachers' skills in interpreting evidence of student learning, *Applied Measurement in Education*, *26*, 191-204, doi:10.1080/08957347 .2013.793185

Sluijsmans, D., Joosten-ten Brinke, D., & Van der Vleuten, C. (2013). *Toetsen met leerwaarde: Een reviewstudie naar de effectieve kenmerken van formatief toetsen* [Testing with learning value: A review study on the effective characteristics of formative assessment].

Spillane, J. P., & Miele, D. B. (2007). Evidence *in* practice: A framing of the terrain. In P. A. Moss (Ed.), *Evidence and decision making* (pp. 46-73). Malden, MA: National Society for the Study of Education.

Staman, L. Visscher, A. J., & Luyten, H. (2014). The effects of professional development on the attitudes, knowledge and skills for data-driven decision making. *Studies in Educational Evaluation, 42*, 79-90. doi:10.1016/j.stueduc.2013.11.002

Stobart, G. (2008). *Testing times: The uses and abuses of assessment.* Abingdon, England: Routledge.

Supovitz, J. (2012). Getting at student understanding – The key to teachers use of test data. *Teachers College Record, 114*(11), 1-25.

Timperley, H. (2009). Using assessment data for improving teaching practice. *Australian College of Educators, 8*(3), 21-27. Retrieved from http://oksowhat.wikispaces.com /file/view/ Using+assessment+data+Helen+Timperley.pdf

Turner, E. O., & Coburn, C. E. (2012). Interventions to promote data use: an introduction. *Teachers College Record, 114*(11), 1-13.

Van der Linden, W. J. (1995). Advances in computer applications. In T. Oakland & R. K. Hambleton (Eds.), *International perspectives on academic assessment* (pp. 105-124). Boston, MA: Kluwer.

Vanhoof, J., Verhaeghe, G., Verhaeghe, J. P., Valcke, M., & Van Petegem, P. (2011). The influence of competences and support on school performance feedback use. *Educational studies, 37*, 141-154. doi:10.1080/03055698.2010.482771

Vermeulen, J. A., & Van der Kleij, F.M. (2012). In Th. J. H. M. Eggen & B. P. Veldkamp (Eds.), *Psychometrics in Practice at RCEC* (pp. 151-179). Enschede: RCEC Cito/University of Twente. doi:10.3990/3.9789036533744.ch13

Weick, K. E. (1969). *The social psychology of organizing*. Reading, MA: Addison-Wesley Pub.

Weick, K. E. (1995). *Sensemaking in organizations*. Thousand Oaks, CA: Sage.

Weiss, C.H. (1998). Improving the use of evaluations; whose job is it anyway? In A. J. Reynolds, & H. J. Walberg (Eds.), *Advances in educational productivity* (pp. 263-276). Greenwich/ London: JAI Press.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*, 3-14. doi:10.1016/j.stueduc.2011.03.001

Wiliam, D., & Leahy, S. (2006). A theoretical foundation for formative assessment. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA.

References Studies Included in the Review and Meta-Analysis

References marked with an asterisk indicate studies included in the meta-analysis

*Bedford, B. (1988). *School effectiveness characteristics and student achievement: a study of relationships in Georgia middle schools*. ERIC document No. EA 020 722.

*Bosker, R. J., & Hofman, A. (1987). Dimensies van schoolkwaliteit: de algemene en milieuspecifieke invloed van scholen op de prestaties en het keuzegedrag van leerlingen [Dimension of school quality. The general and background specific effect of schools]. In J. Scheerens & W. G. R. Stoel (red.), *De effectiviteit van onderwijsorganisaties* (pp. 51-70). Lisse: Swets & Zeitlinger.

Bosker, R. J., Kremers, E. J. J., & Lugthart, E. (1990). School and instruction effects on mathematics achievement. *School Effectiveness and School Improvement*, *1*, 233-248. doi:10.1080/0924345900010401

*Bourke, S. (1986). How smaller is better: some relationships between class size, teaching practices and student achievement. *American Educational Research Journal, 2*3, 558-571. doi:10.3102/00028312023004558

Brandsma, H. (1993). *Basisschoolkenmerken en de kwaliteit van het onderwijs* [Characteristics of primary schools and the quality of education]. (Unpublished doctoral dissertation). Groningen: RION.

*Carpenter S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology 23*, 760-771. doi:10.1002/acp.1507

*Creemers, B., & Kyriakides, L. (2010). School factors explaining achievement on cognitive and affective outcomes: Establishing a dynamic model of educational effectiveness. *Scandinavian Journal of Educational Research 54,* 263-294. doi:10.1080 /00313831003764529

*Clausen, M. (2001). Unterrichtsqualität: eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität [Instructional quality: A question of perspectives?]. Münster: Waxmann.

De Fraine, B., Van Damme, J., Van Landeghem, G., Opdenakker, M. C., & Onghena, P. (2003). The effect of schools and classes on language achievement. *British Educational Research Journal*, *29*, 841-859. doi:10.1080/0141192032000137330

Driessen, G., & Sleegers, P. (2000). Consistency of teaching approach and student achievement: An empirical test. *School Effectiveness and School Improvement*, *11*, 57-79. doi:10.1076/0924-3453(200003)11:1;1-A;FT057

*Gruehn, S. (1995). The compatibility of cognitive and non-cognitive objectives of instruction. *Zeitschrift Für Pädagogik*, *41*(4), 531-553.

Hammond, P., & Yeshanew, T. (2007). The impact of feedback on school performance. *Educational Studies 33*, 99-113. doi:10.1080/03055690601068212

Hill, P. W., & Rowe, K. J. (1998). Modelling student progress in studies of educational effectiveness. *School Effectiveness and School Improvement, 9,* 310-333. doi:10.1080/0924345980090303

*Hofman, R. H. (1993). *Effectief schoolbestuur: een studie naar de bijdrage van schoolbesturen aan de effectiviteit van scholen* [Effective school administration. A study of the school boards' contribution to school effectiveness]. (Unpublished doctoral dissertation). Groningen: Rijksuniversiteit Groningen.

Hofman, R. H., Hofman, W. H. A., & Guldemond, H. (1999). Social and cognitive outcomes: A comparison of contexts of learning. *School Effectiveness and School Improvement, 10*, 352-366. doi:10.1076/sesi.10.3.352.3499

*Klieme, E., & Rakoczy, K. (2003). *Unterrichtsqualität aus Schülerperspektive: Kulturspezifische Profile, regionale Unterschiede und Zusammenhänge mit Effekten von Unterricht*. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider & K.-J. Tillmann (Eds.), *PISA 2000: Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (pp. 334-359). Opladen, Germany: Leske & Budrich.

*Kunter, M. (2004). *Multiple Ziele im Mathematikunterricht* [Multiple goals in mathematics classes]. (Unpublished doctoral dissertation). Berlin: Freie Universität Berlin.

*Kyriakides, L. (2005). Extending the Comprehensive Model of Educational Effectiveness by an Empirical Investigation. *School Effectiveness and School Improvement, 16*, 103-152. doi:10.1080/09243450500113936

*Kyriakides, L., & Creemers, B. P. M. (2008). Using a multidimensional approach to measure the impact of classroom-level factors upon student achievement: a study testing the validity of the dynamic model. *School Effectiveness and School Improvement*, *19*, 183-205. doi:10.1080/09243450802047873

Levačić, R., Steele, F., Smees, R., & Malmberg, L. (2003, September). *The relationship between school climate & head teacher leadership, and pupil attainment: Evidence from a sample of English secondary schools.* Paper presented at the British Educational Research Association Annual Conference, Edinburgh.

*Lockheed, M. E., & Komenan, A. (1989). Teaching quality and student achievement in Africa: the case of Nigeria and Swaziland. *Teaching & Teacher Education*, *5*, 93-115. doi:10.1016/0742-051X(89)90009-7

*Lockheed, M. E., & Longford, N. (1991). School effects on mathematics gain in Thailand. In S. W. Raudenbush, J. D. Willms & J. Douglas (Eds.), *Schools, classrooms and pupils: international studies from a multi-level perspective* (pp. 131-148). San Diego, CA: Academic Press.

Meijnen, G. W., Lagerweij, N. W., & Jong, P. F. (2003). Instruction characteristics and cognitive achievement of young children in elementary schools. *School Effectiveness and School Improvement, 14*, 159-187. doi:10.1076/sesi.14.2.159.14224

Olina, Z., & Sullivan, H.J. (2002). Effects of classroom evaluation strategies on student achievement and attitudes. *Educational Technology Research and Development, 50*, 61-75. doi:10.1007/BF02505025

*Pugh, G., & Telhaj, S. (2003, September). *Attainment effects of school enmeshment with external communities: Community policy, church/religious influence, and TIMSS-R mathematics scores in Flemish secondary schools.* Paper presented at the European Conference on Educational Research, Hamburg.

*Rakoczy, K., E. Klieme, E, Bürgermeister, A., & Harks, B. (2008). The interplay between student evaluation and instruction - Grading and feedback in mathematics classrooms. *Zeitschrift für Psychologie, 216*, 111-124. doi:10.1027/0044-3409.216.2.111

*Reezigt, G. J. (1993). *Effecten van differentiatie op de basisschool*. (Unpublished doctoral dissertation). Groningen: RION.

Reezigt, G. J., Guldemond, H., & Creemers, B. P. M. (1999). Empirical validity for a comprehensive model on educational effectiveness. *School Effectiveness and School Improvement, 10*, 193-216. doi:10.1076/sesi.10.2.193.3503

Rymenans, R., Geudens, V., Coucke, H., Van den Bergh, H., & Daems, F. (1996). *Effectiviteit van Vlaamse secundaire scholen: een onderzoek naar de effecten van het onderwijsaanbod, de tijdsbesteding aan het Nederlands en de schoolkenmerken op de lees- en schrijfprestaties*. Antwerpen: Universiteit van Antwerpen.

Schaub, M., & Baker, D. P. (1991). Solving the math problem: exploring mathematics achievement in Japan and American middle grades. *American Journal of Education*, *99*, 623-642. doi:10.1037/0022-0663.93.2.363

Schildkamp, K. Visscher, A., & Luyten, H. (2009). The effects of the use of a school self-evaluation instrument. *School Effectiveness and School Improvement, 20*, 69-88. doi:10.1080/09243450802605506

Senkbeil, M. (2006). Die Bedeutung schulischer Faktoren für die Kompetenzentwicklung in Mathematik und in den Naturwissenschaften. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, J. Rost & U. Schiefele (Eds.), Pisa 2003: Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres. (pp. 277-308). Münster: Waxmann Verlag GmbH.

*She, H. C., & Fisher, D. (2002). Teacher communication behavior and its association with students' cognitive and attitudinal outcomes in science in Taiwan. *Journal of Research in Science Teaching, 39*, 63-78. doi:10.1002/tea.10009

Van der Grift, W., Houtveen, T., & Vermeulen, C. (1997). Instructional climate in Dutch secondary education. *School Effectiveness and School Improvement, 8*, 449-462. doi:10.1080/0924345970080404

Van der Werf, G., Creemers, B., & Guldemond, H. (2001). Improving Parental Involvement in Primary Education in Indonesia: Implementation, Effects and Costs. *School Effectiveness and School Improvement*, *12*, 447-466. doi:10.1076/sesi.12.4.447.3444

*Vermeulen, C. J. (1987). De effectiviteit van 17 Rotterdamse stimuleringsscholen. *Pedagogische Studiën*, *64*, 49-58.

Willms, J. D., & Somers, M. A. (2001). Family, classroom, and school effects on children's educational outcomes in Latin America. *School Effectiveness and School Improvement, 12*, 409-445. doi:10.1076/sesi.12.4.409.3445

Yelton, B. T., Miller, K., & Ruscoe, G. C. (1994, April). The stability of school effectiveness: comparative path models*. Paper presented at the annual meeting of the AERA, New Orleans.

*Ysseldyke, J., & Bolt, D. M. (2007). Effect of technology-enhanced continuous progress monitoring on math achievement. *School Psychology Review, 36*, 453-467.

*Ysseldyke, J., & Tardrew, S. (2007). Use of a progress monitoring system to enable teachers to differentiate mathematics instruction. *Journal of Applied School Psychology, 24*, 1-28. doi:10.1300/J370v24n01_01

Table A1: Summary of the studies (and samples) on Evaluation at school level

Studies marked with an asterisk indicate studies included in the meta-analysis

Authors (publication year)	Sample	Countries in sample	School type	Measure of evaluation	Outcome measure	Schools (N)	Students (N)	Study design	Statistical technique employed	Prior attainment/ ability included as a covariate	SES included as a covariate	Effect (Fisher z)
Bedford* (1988)		USA	S	Monitoring student progress	Land & Math (mean school score)	70		C	P	N	N	-.030
Bosker & Hofman* (1987)		Netherlands	P	Analysis and use of pupil data for didactical help of students	Lang & Math	72	356	C	ML	Y	Y	-.065
Brandsma (1993)		Netherlands	P	Evaluation pupil progress and program School process evaluation	Lang & Math	208		C	ML	Y	Y	
Creemers & Kyriakides* (2010)		Cyprus	P	School process evaluation	Lang, Math & Other	50	2503	C	ML			.062
Hammond & Yeshanew (2007)		UK	S	School performance feedback (paid vs unpaid)	Lang & Math		4093	E	ML			
Hofman* (1993)		Netherlands	P	Evaluation pupil progress and school process	Lang & Math	220	4000	C	P	N	N	.214
Hofman et al. (1999)		Netherlands	P	Evaluation of education at the school level The degree of pupil evaluation in school to acquire consistency in the learning process	Math	103		C	ML	Y	Y	

Authors (publication year)	Sample	Countries in sample	School type	Measure of evaluation	Outcome measure	Schools (N)	Students (N)	Study design	Statistical technique employed	Prior attainment/ ability included as a covariate	SES included as a covariate	Effect (Fisher z)
Kyriakides* (2005)		Cyprus	P	Assessment system focused on formative purposes	Lang & Math	81	1721	C	ML	Y	Y	.066
Reezigt et al. (1999)		Netherlands	P	Evaluation policy: pupil monitoring system, discussion of results in team		127		C	ML	Y	Y	
Schildkamp et al. (2009)		Netherlands	P	Conceptual and instrumental use of school self-evaluation	Lang & Math	79		C	ML	Y	Y	
Senkbeil (2006)		Germany	S	Evaluation practices	Math & Other	141		C	ML	Y	Y	
Van der Grift et al. (1997)		Netherlands	S	Frequent monitoring of student results Acting upon results of tests	Math	100		C	ML	Y	Y	
Vermeulen* (1987)		Netherlands	P	Monitoring pupil progress on a regular basis	GAA (mean school score)	17		C	Part	N	N	.255
Van der Werf et al. (1999)		Indonesia	P	Evaluation of school quality	Math	81		C	ML	Y	Y	
Yelton et al.* (1994)		US	P	Use of test data for assessing school wide performance and planning for improvement	Lang & Math (mean school score)	49		C	P	N	N	.273

Notes: * included in quantitative meta-analysis, P: primary education; S: secondary education; Lang: language; GAA = general academic achievement; E: Experimental design; Q: Quasi-experimental design; C: Correlational design; P: Pearson correlation analysis, Part: Partial correlation, R: Regression analysis; ML: Multilevel analysis, PA = path analysis, SEM: Structural Equation Modeling

Table A2: Summary of the studies (and samples) on Evaluation at class level
Studies marked with an asterisk indicate studies included in the meta-analysis

Authors (publication year)	Sample	Countries in sample	School type	Measure of evaluation	Outcome measure	Schools (N)	Students (N)	Study design	Statistical technique employed	Prior attainment/ ability included as a covariate	SES included as a covariate	Effect (Fisher z)
Bourke* (1986)		Australia	P	Teacher monitoring student work	Math (mean class score)	63		C	P	N	N	.020
									Path	Y	N	
Brandsma (1993)		Netherlands	P	Performance feedback	Lang & Math	208		C	ML	Y	Y	
Carpenter et al.* (2009)		US	S	Timing of feedback	Other	5	62	E	A	N	N	.236
Clausen* (2001)		Germany	S	Monitoring Individual reference norm orientation	Math (mean class score)	53		C	P	N	N	-.020
De Fraine et al. (2003)		Belgium	S	Feedback	Language	111	1834	C	ML	Y	Y	
Driessen & Sleegers (2000)		Netherlands	P	Registration pupil progress Purposes of assessment	Lang & Math	492	7410	C	ML	N	Y	
Gruehn* (1995)		Germany	S	Monitoring Individual reference norm orientation Diagnostic competences	Math (mean class score)	137		C	P	N	N	.147
Hill & Rowe (1998)		Australia	S	Feedback Teacher feedback	Language	365	6423	C	ML	Y	Y	
Hofman et al. (1999)		Netherlands	P	Feedback to pupils Diagnostic practice for pupils with learning problems	Math	103	2023	C	ML	Y	Y	

Authors (publication year)	Sample	Countries in sample	School type	Measure of evaluation	Outcome measure	Schools (N)	Students (N)	Study design	Statistical technique employed	Prior attainment/ ability included as a covariate	SES included as a covariate	Effect (Fisher z)
Klieme & Rakoczy* (2003)		Germany	S	Individual reference norm orientation	Math (mean class score)	408		C	P	N	N	.070
Kunter* (2004)		Germany	S	Individual reference norm orientation and diagnostic competences Feedback	Math	80	1900	C	ML	Y	N	-.045
Kyriakides* (2005)		Cyprus	P	Providing feedback	Lang & Math	81	1721	C	ML	Y	Y	.094
Kyriakides & Creemers* (2008)		Cyprus	P	Feedback quality: type of feedback that the teacher gives to the students and the way students use the teacher feedback	Lang, Math & Other	108	2503	C	ML	Y	Y	.060
Levacic et al. (2003)		UK	P	Monitoring and reward Monitoring and identifying underachievers Positive reinforcement	Language (Key stage 3 and GSCE)	124	870	C	ML	Y	Y	
Lockheed & Longford* (1991)		Thailand	S	Providing feedback to students	Math	60	2076	C	ML	N	Y	.095
Rakoczy et al.* (2008)		Germany	S	Evaluative feedback (correct) Evaluative feedback (incorrect) Informational feedback	Math	10	240	Q	R	Y	N	.007

Authors (publication year)	Sample	Countries in sample	School type	Measure of evaluation	Outcome measure	Schools (N)	Students (N)	Study design	Statistical technique employed	Prior attainment/ ability included as a covariate	SES included as a covariate	Effect (Fisher z)
Reezigt* (1993)		Netherlands	P	Monitoring and feedback	Lang & Math	218	4369	C	R	Y	Y	.023
Reezigt et al. (1999)		Netherlands	P	Monitoring Feedback Corrective instruction	Language	258	1531	C	ML	Y	Y	
Ryemenans et al. (1996)		Belgium	S	Purposes of assessment	Language	24		C	ML	Y	Y	
Senkbeil (2006)		Germany	S	Use of evaluation results	Math & Science	144		C	ML	Y	Y	
She & Fisher* (2002)		Taiwan	S	Encouragement & praise	Science	28	1138	C	R	N	N	-.125
Van der Grift et al. (1997)		Netherlands	S	Monitoring Acting upon student results	Math	109	2938	C	ML	Y	Y	
Van der Werf (2001)		Indonesia	P	Monitoring work Use of test results	Lang, Math & Science	81	1854	C	ML	Y	Y	
Ysseldyke & Bolt* (2007)		US	P&S	Progress monitoring and instructional management system	Math	80	1274	Q	A	Y	N	.192
Ysseldyke & Tardrew* (2007)	P	US	P	Progress monitoring and instructional management system	Math	21	328	Q	A	Y	N	.191
	S	US	S	Progress monitoring and instructional management system	Math	21	46	Q	A	Y	N	.191

Notes: * included in quantitative meta-analysis, P: primary education; S: secondary education; Lang: language; GAA: general academic achievement; E: Experimental design; Q: Quasi-experimental design; C: Correlational design; A: ANOVA, P: Pearson correlation analysis, R: Regression analysis; ML: Multilevel analysis, PA = path analysis, SEM: Structural Equation Modeling

Table A3: Summary of the studies (and samples) on Assessment

Studies marked with an asterisk indicate studies included in the meta-analysis

Authors (publication year)	Sample	Countries in sample	School type	Measure of evaluation	Outcome measure	Schools (N)	Students (N)	Study design	Statistical technique employed	Prior attainment/ ability included as a covariate	SES included as a covariate	Effect (Fisher z)
Bourke* (1986)		Australia	P	Use of oral tests for assessment	Math (mean class score)	63		C	P	N	N	-.172
Bosker, Kremers & Lugthart (1990)		Netherlands	P	Use of evaluative tests	Math	44		C	ML	N	Y	
Carpenter et al.* (2009)		US	S	Review method: Test/study versus study Test/study versus no review	Other	5	62	E	A	N	N	.206
Driessen & Sleegers (2000)		Netherlands	P	Frequency of tests	Lang & Math	492	7410	C	ML	N	Y	
Kyriakides & Creemers* (2008)		Cyprus	P	Assessment	Lang, Math & Other	108	2503	C	ML	Y	Y	.015
Lockheed & Komenan* (1989)	Nigeria	Nigeria	S	Weekly minutes for testing and grading	Math	41	700	C	R	N	Y	.026
	Swaziland	Swaziland	S	Weekly minutes for testing and grading	Math	25	587	C	R	N	Y	-.079
Meijnen et al. (2003)		Netherlands	P	Number of evaluation procedures	Language	42	282	C	ML	Y	Y	
Olina & Sullivan (2002)		Latvia	S	Teacher evaluation vs no evaluation	Other		189	Q	A	N	N	
Pugh & Telhaj* (2003)		Belgium	S	Teachers time spent by teacher on scrutiny tests/exam	Math		5259	C	R	Y	Y	.028

Authors (publication year)	Sample	Countries in sample	School type	Measure of evaluation	Outcome measure	Schools (N)	Students (N)	Study design	Statistical technique employed	Prior attainment/ ability included as a covariate	SES included as a covariate	Effect (Fisher z)
Reezigt* (1993)		Netherlands	P	Frequency of testing Frequency of diagnostic testing	Lang & Math	205	4369	C	R	Y	Y	.001
Reezigt et al. (1999)		Netherlands	P	Use of curriculum tests	Language	258	1531	C	ML	Y	Y	
Rymenans et al. (1996)		Belgium	S	Use of curriculum embedded tests Number of tests for language	Language	24		C	ML	Y	Y	
Schaub & Baker (1991)		US & Japan	S	Number of tests	Math (mean class score)	430		C	R	N	N	
Van der Werf et al. (2001)		Indonesia	P	Frequency of testing	Lang, Math & Science	81	1854	C	ML	Y	Y	
Willms & Somers (2001)	11 countries	11 Latin American countries	P	Pupils are tested	Language		100 (average)	C	ML	Y	Y	

Notes: * included in quantitative meta-analysis, P: primary education; S: secondary education; PS: Primary and Secondary education; Lang: language; GAA: general academic achievement; E: Experimental design; Q: Quasi-experimental design: C: Correlational design; A: ANOVA, P: Pearson correlation analysis, R: Regression analysis; ML: Multilevel analysis, PA = path analysis, SEM: Structural Equation Modeling

5

Time effects in education; A meta-analysis[1]

[1] This chapter is based on:
Hendriks, M. A., Luyten, J. W., Scheerens, J., & Sleegers, P. J. C. (2014). Meta-analyses. In J. Scheerens (Ed.), *Effectiveness of time investments in education* (SpringerBriefs in Education) (pp. 55-142). Cham: Springer. doi:10.1007/978-3-319-00924-7_2

Abstract

Previous meta-analyses reported small to moderate effects of time for schooling and teaching and homework on student achievement. The present study updates the research available up to 2005 and considers both the general overall as well as the differential effects of facets of learning time and homework. The meta-analysis included 12 studies on learning time in schools, and 23 studies for homework. Analyses (using random effects models) revealed small but significant overall effects for time in schools and homework at both individual and class level (r = .046, r = .044 and r = .058 respectively) as well as for two of the nine facets of time. The effects found are lower than those reported in most previous meta-analyses. Tentative explanations for these discrepancies are discussed and suggestions for further research are made.

Introduction

Relevance of Time in Education

Time for schooling and teaching is considered one of the key variables to improve educational outcomes and the quality of schooling (see e.g. Scheerens, 2014). The underlying notion, namely that good schooling and teaching depends on the "exposure" of students is clear and plausible. As a consequence, national and local politicians, educational policy-makers and practitioners in many countries are involved in developing strategies to expand time in schools, ranging from expanding the school year, school week or school day, instructional time, home work and home support. The nature and the effectiveness of these strategies have been discussed by policy makers and researchers over the last decades.

In order to design evidence-based policy strategies to foster effective teaching and schooling, systematic knowledge about the extent to which time affects students' outcomes is essential. Several researchers have conducted meta-analyses on the effect of learning time in school and homework on student achievement (see e.g. Cooper, 1989; Cooper, Robinson & Patall, 2006; Fraser, Walberg, Welch & Hattie, 1987). Although this research has provided more insights into the effectiveness of time, it has its limitations. A broad range of different operational definitions of time, ranging from "statutory", official school or teaching hours, time on task, "quality time" to the amount and frequency of homework at individual and school level, was used in the different studies on which the separate meta-analyses were based. As the effects of this mixture of different specifications were thrown together in the meta-analyses, the findings can only be interpreted as a general overall effect of time. They do not inform us about the effectiveness of specific facets of learning time and homework, however. Furthermore, because of methodological flaws in the original studies as well as the meta-analyses, the effect-sizes varied considerably and should be interpreted with caution (Canadian Council on Learning, 2009; Kohn, 2006; Seidel & Shavelson, 2007; Trautwein, Lüdtke, Schnyder & Niggli, 2006). E.g. Kohn (2006) provides a "taxonomy of abuses" in studies that examined the effect of homework and came to the conclusion that there is virtually no evidence that unequivocally supports the expectation that homework has beneficial effects on academic achievement or on attitudes that would be supportive of independent learning.

Finally, most of the meta-analyses are based on empirical studies published before 2005. The meta-analysis reported in this article contributes to this line of research in three ways: by taking different facets of learning time and homework into account, by meeting the methodological challenges as discussed in the critical reviews of earlier work and by providing a more up to date picture on the effectiveness of time for schooling and teaching and homework. In doing so, this study attempts to increase our understanding of the role of time as a factor in educational productivity.

Conceptualization of Time as a Factor in Educational Productivity

In conceptualizing and assessing the effects of time on student outcomes, at least four issues should be addressed. First of all time can be defined in a "gross" and "net" way. The officially mandatory school time and lesson time per subject, usually indicated as "allocated time", is to be seen as a gross measure. The time schools and teachers actually realize, sometimes indicated as the "exposed time", is often considered a good indicator of "net time". Closer to the real "net (teaching) time" that students are exposed to is the proportion of time that remains of a lesson after subtraction of the time a teacher needs for classroom management. Stallings and Mohlman (1981) estimate this latter percentage (time for classroom management) at 15% while Lam (1996) estimated this proportion at 7%, based on the analysis of logs. Ultimately the effective time students are engaged in learning could be defined as the percentage of on-task behavior of students during lessons; often referred to as "time on task".

Secondly, the issue of educational time does not remain limited to optimizing regular "within school time". Since decades, policies to expand the school year, school week or school day are applied in countries like the USA, Japan and Korea, and more recently such policies also happen in the Netherlands (Oberon, 2009). Homework and homework support can be placed as an in-between category, on the one hand as closely linked to regular within school teaching, on the other hand as added, out of school time.

A third issue that needs to be addressed is the nature of the relationship between time facets and educational achievement. Research has shown that the estimated positive effect of time on student outcomes, the more time the better performance, is not linear, and shows diminishing returns (Keith, 1982 as cited in Keith, Diamond-Hallam & Fine, 2004). This means that after a certain level the incremental benefits of even more time become smaller.

A fourth and final issue is related to the relation between quantity and quality of education. The assumption that more effectively used time will enhance student performance, implicitly suggests that the additional time is well used in terms of covering content and offering good instruction. In research this dependency presents the challenge to unravel time, content and instructional process influences (see e.g. Cool & Keith, 1991). One might even say that quality and time, or "quantity and quality" of education, to use

Walberg's words (Walberg, 1986) provide a trade-off, in the sense that high quality education can, to some degree, compensate for long lesson hours. Finland's impressive achievement on international assessments, such as TIMSS and PISA, may be considered as an illustrative case in point.

State of the Art: Results from Previous Meta-Analyses on Learning Time in Schools and Homework

Learning time in schools

The impact of learning time in school has been analyzed in three more recent meta-analyses (Kyriakides, Creemers, Antoniou & Demetriou, 2010; Scheerens, Luyten, Steen & Luyten-de Thouars, 2007; Seidel & Shavelson, 2007). These meta-analyses all focused on a broad conceptualization of time and examined the effect sizes of time on student outcomes, as well as those of other school and instruction effectiveness enhancing variables. Effect sizes were expressed either as a standardized mean difference between an experimental and a control group (indicated with coefficient Cohen's *d*) or as correlations (indicated with the coefficient *r,* expressing the product moment correlation. Standardized mean differences and correlation coefficients (*r* and *d*) are convertible to one another[1]. Below, in subsequent tables, effect sizes will be expressed in correlations.

Table 5.1 provides an overview of the three recent meta-analyses that analyzed the impact of time on student achievement. The average effect size across the three meta-analyses is r = .11.

Table 5.1
Overview of earlier meta-analyses on the effects of learning time on student achievement

Authors	Conceptualization of time of studies included	Estimated mean effect size	Number of studies included
Scheerens et al. (2007)	Learning time	r =.15	30
Seidel & Shavelson (2007)	Learning time, opportunity to learn and homework	r =.03	34
Kyriakides et al. (2010)	School policy on the quantity of teaching	r =.16	18

In the meta-analysis of Scheerens et al. (2007) a range of operational definitions of learning time was used in the studies, varying from time at school and time at classroom level to monitoring of absenteeism and classroom management. The estimated mean effect size the authors reported was r = 0.15. Moderator analyses showed that studies conducted in secondary schools yielded significantly lower effect sizes than studies employed in primary

[1] Converting from r to d (Borenstein et al., 2009, p. 48), is as follows:

$$\mathrm{d} = \frac{2\mathrm{r}}{\sqrt{1-\mathrm{r}^2}}$$

schools (r= 0.01 in secondary schools versus r = 0.38 in primary schools). Furthermore, in studies in which multi-level analyses were used, significantly lower effect sizes were found than in studies where researchers did not conduct these analyses (r = -0.02 versus r = 0.40 respectively). Compared to other school effectiveness enhancing variables in this meta-analysis, time had still the highest effect on student achievement together with curriculum quality/opportunity to learn (r = .15), achievement orientation (r = .14) and orderly climate (r =. 13).

Kyriakides et al. (2010) conducted a meta-analysis, based on studies that involved school and classroom level variables as part of the dynamic model of educational effectiveness. School policy on the quantity of teaching was one of the school level variables included and covered facets of the school policy of time, like policy on the management of teaching time, policy on teacher and student absenteeism, policy on homework and policy on lesson schedule and time table. The effect size they found for school policy on the quantity of teaching (r = .16) was comparable to effect sizes the authors reported for other school level variables including opportunity to learn (r = .15), quality of teaching (r = .17) and student assessment (r =.18).

The meta-analysis of Seidel & Shavelson (2007) focused on teaching effects of time on student outcomes. Time for learning in this meta-analysis covered both time on task, opportunity to learn and homework. Seidel & Shavelson included three types of outcome measures: learning processes, motivational-affective outcomes and cognitive outcomes. Concerning the association with cognitive outcomes, they found a small effect size for learning time (r = 0.03) for learning time), while for learning processes or motivational–affective outcomes the effect sizes found were considerably higher (r = .14 and r = .12 respectively). However, the number of studies that included learning processes or motivational–affective as outcome measures were considerably lower than those that focused on cognitive outcomes.

Although these meta-analyses have increased our insights into the effects of learning time in schools on student outcomes, there are several limitations. An aspect already mentioned is the broad range of different operational definitions of time as used in the studies on which the separate meta-analyses were based Moreover, studies included in the meta-analyses do not always control for other effectiveness enhancing variables, such as e.g. opportunity to learn and quality of instruction. Learning time in schools or homework correlates with achievement when studied in isolation but because of multicollinearity the effects are likely to decrease or even disappear when other effectiveness enhancing variables are "controlled for". Furthermore, with the exception of the meta-analysis from Seidel & Shavelson, there is no clear indication that previous meta-analyses were based on studies that had controlled for student prerequisites, and this might explain the differences in mean effect sizes reported. A final limitation of most of the earlier meta-analyses is that effects of time for different subjects and different types of students are not examined. On the basis of a review of the literature, The Core Academic Learning Time Group (2002) states that the effect of time is stronger for highly structured subjects, like mathematics, science and reading, than for the more open subjects like art and social studies. There is also a strong suggestion in the literature that sufficient time is especially important for the weaker students (Keith, 1982, The Core Academic Learning Time Group, 2002).

Homework

The relationship between homework and student achievement was examined in four meta-analyses (Cooper, 1989; Cooper, Robinson & Patall, 2006; Paschal, Weinstein & Walberg, 1984; Scheerens et al., 2007) and concerned meta-analyses either synthesizing studies with a (quasi-) experimental research design examining the impact of homework versus no homework or meta-analyses examining bivariate associations between homework and achievement (see Table 5.2). The average effect size across these six meta-analyses amounts to r = .16.

Table 5.2
Overview of earlier meta-analyses on the effects of homework on student achievement

Authors	Conceptualization of homework in studies included	Estimated mean effect size	Number of studies included
Paschal, Weinstein & Walberg (1984)	Homework vs No Homework	r = .18.	15
Cooper (1989)	Homework vs No Homework	r= .10	17
Cooper (1989)	Homework	r= .19	50
Cooper et al. (2006)	Homework vs No Homework	r= .29	6
Cooper et al. (2006)	Homework	r = .24	32
Scheerens et al. (2007)	Homework	r= .07	21

Cooper's synthesis (1989) on the effect of homework on student outcomes was based on nearly 120 studies and included three types of evidence, two of them relevant within the context of our study. The first type of study focused on (quasi-) experimental studies examining the impact of homework versus no homework. For these studies Cooper found a moderating influence of grade level: the effect found for high schools students was twice the magnitude of the effect found in junior high schools. In a second type of evidence Cooper (1989) examined correlational studies into the bivariate relation between time spent on homework and achievement. The results showed a positive relationship between homework and achievement in 43 studies, while in seven studies an average negative association was found. An interaction effect with grade level was found as well: Effects of time spent on homework were small in primary schools and middle schools while for high schools a substantial effect was were found (r =.25).

Cooper's (1989) results were criticized by Trautwein and Köller (2003) for methodological shortcomings. In especially Trautwein and Köller pointed to the impact of study characteristics on the magnitude of the effect sizes found. Of the studies included in the meta-analysis, only four studies controlled for student background characteristics or reported gains. In these studies a negative small effect size was found (r = -.04) suggesting that the positive overall effect might be confounded by the effect of student background characteristics.

In order to provide a more up to date picture on the effectiveness of time for homework and to meet the methodological critiques on their earlier meta-analysis Cooper et al. (2006) conducted a new synthesis on homework. The 68 studies included in this meta-analysis were categorized into three types of research design: i.e. experimental studies in which homework and no-homework conditions were manipulated, studies using cross sectional data investigating multivariate correlations, and studies that examined bivariate correlations.

In the cross-sectional studies multivariate analysis techniques (multiple regression, path analysis or structural equation modeling) were conducted to investigate the relationship between time spent on homework and student achievement. The reported standardized β-weights ranged from .05 to .28 and effects found were quite similar across subject domains. The majority of studies was conducted in higher education and based on longitudinal data from large US national surveys. Cooper et al. also investigated the role of confounding variables but could not draw any conclusions regarding their impact as: "the number and type of predictors in each model was complex, varied considerably from model to model, and potentially were confounded with one another across studies" (Cooper et al., 2006, p. 48).

In the studies examining bivariate correlations between time spent on homework and student achievement, 50 were in the positive direction and 19 negative.

Scheerens et al. (2007) found a small average effect size of r = .07. Using moderator analyses, it appeared that the effect sizes for homework were substantially higher in the USA and the Netherlands, than in all other countries. No effect of subject domain was found.

The findings of these four meta-analyses showed small to medium positive effects of homework on student achievements. As was the case with the meta-analyses on learning time in school, most of the studies on which the meta-analyses on homework were based used a mixture of different specifications of homework. This means that the reported effects just indicate a general overall effect of homework. This is an important limitation as the findings of individual studies (De Jong, Westerhof & Creemers, 2000; Trautwein, 2007; Trautwein et al., 2006), have indicated that it matters a lot, which operational measure of homework is used and at which analytical level homework is examined For example*, time spent on homework* has shown to have mixed effects. When multilevel modeling is applied, effects show up only at the aggregate (school or classroom) level while effects are negligible or negative at individual student level (Trautwein, 2007; Trautwein & Köller, 2003; Trautwein, Schnyder, Niggli, Neumann & Lüdtke, 2009). Furthermore, studies (De Jong et al, 2000; Mooring, 2004) have shown that *amount of homework,* defined as the quantity of content covered during homework assignment, had strong positive effects on student outcomes. In addition, Trautwein (2007) reported medium to strong effect sizes of *homework effort*, based on students' ratings of the effort invested in homework, Homework effort and amount of content covered in homework assignments appear to be more powerful associates of achievement than time spent on homework and frequency of homework assignments.

The Present Study

As illustrated in the above, meta-analyses examining the effects of regular school time usually throw together a range of different "treatments", varying from increments in "statutory", official school or teaching hours, to more efficient use of teaching time, time on task, and "quality time" (see Scheerens et al., 2007; Seidel & Shavelson, 2007). Moreover, in order to be effective it is obvious that time should be "filled" with relevant content and effective instruction. In empirical studies on which the meta-analyses are based, these variables are not always controlled for, so that "time" effects may also include the effects of content covered and quality of instruction. Individual studies on the effects of homework seem to underline this point. On the few occasions that pure time effects, in terms of frequency and duration of homework assignments could be separated from content covered, it was the latter facet, indicated as "amount" of homework, which appeared to be the most important (De Jong et al., 2000). Due to these limitations, the coefficients reported in the meta-analyses should be interpreted with some caution. A second reason to interpret the coefficients carefully has to do with methodological flaws in the original studies as well as the meta-analyses (Canadian Council on Learning, 2009; Kohn, 2006; Seidel & Shavelson, 2007; Trautwein et al., 2006). Despite these limitations, extra time should be seen as an important condition to "increase well targeted exposure to content" as a relatively strong mediator of student achievement in comparison to other educational effectiveness enhancing conditions.

By conducting a new meta-analysis, we attempt to validate earlier findings with findings from recent studies and provide an up to date picture on the effectiveness of learning time in schools and homework. In order to address the limitations of earlier meta-analyses, we will focus on both the effects of specific facets of learning time and homework and general overall effects, and we will try and meet some of the methodological challenges as discussed in the above.

Method

Learning Time in Schools and Homework: Definition and Facets

In the case of learning time in schools, a distinction was made between allocated time, instruction time and time on task. Allocated time refers to official teaching hours. Instruction time refers to the part of the allocated time that is spent on instruction, where the difference between allocated and "net" instruction time may be caused by the time the teachers need for classroom management. Time on task is defined on the basis of engaged student behavior, being the time a student manifests on-task behavior.

Three types of measures of homework were distinguished, homework frequency, homework time and amount of homework. The first refers to the number of times students get homework assignments, the second to the time they spent on homework, while amount of homework refers to the amount of subject matter that the students covered during homework. Moreover, also a distinction is made between homework as measured at the individual student level and homework measured at the school or classroom level.

The meta-analyses reported in this article consists of a reanalysis and extension of earlier meta-analyses published by Scheerens and Bosker (1997), Scheerens, Seidel and others (2005) and Scheerens, Luyten, Steen and Luyten-de Thouars (2007). The extension consisted of studies that were published in the period 2005-2011.

Search Srategy and Selection Criteria

To select studies on learning time in school and homework published in the period 2005-2011 a computer assisted search was conducted in November 2011.

The following online databases were used: Web of science (www.isiknowledge.com); Scopus (www.scopus.com); and ERIC and Psycinfo (provided through Ebscohost). The databases were primarily explored using the same key terms as used in the meta-analysis by Scheerens et al. (2007): school effectiveness, educational effectiveness, teacher effectiveness, effective teaching, effective instruction, instruction, mastery learning, constructivist teaching, mathematics instruction, reading instruction, science instruction, mathematics teaching, reading teaching, science teaching. Each effectiveness keyword was crossed with each of the following output keywords: value added, attainment, achievement, learn* result*, learn* outcome*, learn* gain, student* progress and with each the time variables of interest for this meta-analysis: learning time at school and homework). A total of 13047 publications matched combinations of the keyword. After removing the duplicate publications 10626 unique publications were selected for the next step.

The next step then was to examine the title and abstract of each publication to determine whether the study met the following in- and exclusion criteria:

- The study had to include an independent variable measuring learning time at school or time spent on homework at student, class or school level.
- The study had to include a measure of cognitive student achievement in mathematics, language, science or other school subjects as the dependent variable. Examples include scores on standardized tests, achievement gain scores and grades in subject areas.
- The study had to focus on primary or secondary education (for students aged 6-18). Studies that focused on preschool, kindergarten or on post-secondary or tertiary education were excluded.
- The study had to focus on regular students. Studies containing specific samples of students in regular schools (such as students with learning, physical, emotional, or behavioral disabilities) or studies conducted in schools for special education were excluded from the meta-analysis.
- The study had to be published or reported between 2005 and 2011. Studies published as online first publication in 2011 were also included.
- The study had to be written in English, German or Dutch.

- The study had to have estimated the relationship between a measure of learning time at school or homework time and student achievement. This means that the study had to provide one or more effect sizes or had to include sufficient quantitative information to permit the calculation of an effect size.

If the abstract of the publication did not include sufficient information to decide that the publication met the in- or exclusion criteria, the full text of the publication was reviewed by one of the researchers. In total 382 publications passed to the second round for full-text review. In addition, to identify additional published studies, recent reviews and books on learning time at school, homework and out-of-school learning time were examined, as well as the literature review sections from the obtained articles, chapters, research reports, conference papers and dissertations.

The review of full text publications resulted in 30 publications covering the period 1985-2011 to be coded or rechecked in the coding phase.

Coding Procedure

Lipsey and Wilson (2001) have defined two levels at which the data of the study should be coded: the study level and the level of an effect size estimate. According to the authors a study can be defined as “a set of data collected under a single research plan from a designated sample of respondents” (Lipsey & Wilson, p. 76). A study may contain different samples, when the same research is conducted on different samples of participants (e.g. when students are sampled in different grades, cohorts of students or students in different stages of schooling - primary or secondary-) or when students are sampled in different countries. An estimate is an effect size, calculated for a quantitative relationship between an independent and dependent variable. As a study may include, for example, different measurements of the *independent* variable (such as allocated learning time and time on task in the case of learning time in schools), different achievement measures to measure the *dependent* variable (such as different subtests covering different domains of subject matter), multiple assessments of pupils at several time-points, and different statistical analyses (e.g. Pearson correlation and regression), a study may yield many effect sizes, each estimate different from the others with regard to some of its details.

The studies selected between 2005 and 2010 were coded by the researchers applying the same coding procedure as used by Scheerens et al. (2007). The coding form included five different sections:

- *Report and study identification.*
  This section recorded the author(s), the title and the year of the publication;
- *Characteristics of the independent (time) variable(s) measured.*
  In this section the conceptualization of the time variable(s) used in the study (i.e. learning time at school, homework time at pupil level, homework time at class/school level) as well as the subcategories or types of the time variables distinguished (allocated time, instruction time and time on task for learning time at school and homework frequency, homework

time and amount of homework for homework at individual level and homework at class/school level respectively) were coded. The operational definitions of the time variables used in the studies were recorded too.

- *Sample characteristics.*
  The sample characteristics section recorded the study setting and participants. For study setting the country or countries included in the study were coded. With regard to participants, the stage of schooling (primary or secondary level) the sample referred to was coded as well as the grade or age level(s) of the students the sample focused on. The number of schools, classes and students included in the sample were recorded as well.
- *Study characteristics.*
  In this section the research design chosen, the type of instruments used to measure the time variable(s), the statistical techniques conducted and the model specification were coded. For research design we coded whether the study applied a quasi-experimental or experimental design and whether or not a correlational survey design was used. With regard to the type of instruments used we coded whether a survey instrument or log was used, the respondents (students, teachers, principals and/or students), and whether data were collected by means of classroom observation or video-analysis or (quasi-)experimental manipulation. The studies were further categorized according to the statistical techniques conducted to investigate the association between time and achievement. The following main categories were employed: ANOVA, Pearson correlation analysis, regression analysis, path analysis/LISREL/SEM and multi-level analysis. We also coded whether the study accounted for covariates at the student level, i.e. if the study controlled for prior achievement, ability and/or student social background. For learning time at school we coded whether, in addition to the time variable used, (other) process variables at class or school level were included in the study as well.
- *Time effects (effect sizes).*
  Finally, the time effects section recorded the effects sizes, either taken directly from the selected publications or calculated (see section calculation of effects sizes below). The effect sizes were coded as reflecting the types of outcome variables used (i.e. achievement test score, value-added output indicator, gain score, attainment measure, grade) as well the academic subject(s) addressed in the achievement measure. Four groups of subjects were distinguished in the coding: language, mathematics, science and other subjects.

Calculation of Effect Sizes

In the majority of studies that were fully coded in our database, coefficients from regression and multilevel analysis were reported. Standardized regression coefficients were substituted directly for correlation coefficients as coefficients from multiple regression correspond to *r* equally well (for β coefficients between -.50 and .50, see Peterson and Brown, 2005). For studies that reported unstandardized coefficients, standardized coefficients were computed if the standard deviations of the explanatory variable and the achievement measure were reported in the publication. This was only possible for a minor number of studies. In these cases we applied the formulae presented in Hox (1995, p. 25) to calculate the standardized regression coefficient and standard error.

For the majority of studies that reported unstandardized regression coefficients, we were not able to calculate standardized coefficients. Therefore these studies were excluded from the quantitative meta-analysis. However, to not throw away the information from these studies we also used a vote counting procedure (Bushman & Wang, 2009), which comes down to counting the number of positive significant, negative significant and non-significant associations in all studies for each of the three main time variables: learning time in schools, homework time at pupil level, homework time at class level. We used a significance level of α=.05. The results are reported in Hendriks, Luyten, Scheerens and Sleegers (2014).

In some studies multiple techniques for data-analysis were applied, e.g. bivariate Pearson correlations and regression or multilevel analysis. For these studies the coefficients of the most appropriate method (regression or multilevel) were included in the meta-analysis. For studies in which bivariate or partial correlations were used only or for studies for which we were not able to calculate standardized regression coefficients, the estimated Pearson correlation coefficients were included in the meta-analysis. For studies that applied regression or multilevel modeling and in which different (intermediate and final) models were presented, the coefficient(s) from the most fully identified model without interaction effects were used for the meta-analysis.

The unit of analysis for this meta-analysis was the independent sample. Some studies however reported multiple effect size estimates for different analyses examining the association between a measure of time or homework and achievement in the same sample. For example, when a study used two different measurements of the homework variable (e.g., time spent on homework and frequency of homework) and also assessed the impact of each homework variable on two outcome measures (e.g. Dutch an English language achievement), then this study yields four effect sizes. As these effect sizes cannot to be assumed statistically independent (see Bennett, 2011, Cooper et al., 2009, Lipsey & Wilson, 2001), these multiple effect sizes were averaged to yield a single mean effect size at sample level.

Average effect sizes were computed when:

- Multiple measures or operationalizations of the same explanatory variable were included in the same analysis (e.g. homework measured both by a teacher questionnaire and a student questionnaire or homework time and homework frequency);
- Multiple measures of the dependent variable were used to assess student achievement (e.g. when both a reading and writing test are used to measure language achievement or when achievement tests are used in different subjects, e.g. language and math);
- Achievement was measured at different times in the same sample: e.g. at the end of year1, year 2, year 3 and year 4 as was the case in the longitudinal study by Kyriakides & Creemers (2008).

Effect sizes were not averaged in the following cases:

- Analyses were performed per country in case more countries were included in a study (e.g. Japan and the United States).
- Different school levels were included (e.g. both primary and secondary level).
- Different grade levels from the same school level were included in the analysis (e.g. both grade 4 and 6 in primary school).

The final database included 16 samples for learning time in schools, 19 samples for homework at pupil level and 12 samples for homework at class level. Tables A1, A2 and A3 in the Annex provide summaries of the studies included in the meta-analyses.

In order to compare the different effects size estimates used in the studies, we transformed the reported effects size estimates into Fisher's Z using the formulae as presented by Lipsey and Wilson (2001). Fisher's Z was thus used as the effect size estimate for our meta-analysis to analyze the effects of learning time in school and homework on student outcomes.

Weighing of Effect Sizes

To calculate average effect sizes weighted and unweighted procedures can be used. In the unweighted procedure, each effect size is given an equal weight in calculating the average effect size.

In the weighted procedure the weights used can be based on a fixed effects model or random effects model. In a fixed model it is assumed that the true effect size is the same in all samples included in the meta-analysis and that the random influence on the effect sizes stems from sampling error alone. In the random effects model, because of real study-related differences (such as variations in study designs settings, measurements of the independent variable, model specifications etc.), the true effect size is expected to be similar but not identical across samples. In the random effects model the variance between effect sizes is thus due to the within sample variance (like the fixed effects model) plus the between-sample variance (variance randomly distributed across samples (Borenstein, Hedges, Higgings & Rothstein, 2009; Lipsey & Wilson, 2001).

In our meta-analysis a random effects model is considered most appropriate because of large differences in settings, designs, measurements instruments and statistical techniques used in the different studies. Each estimate is weighted by both the inverse of its' within sample variance and the estimate of the between-samples variance (Borenstein et al., 2009).

Data Analysis

Overall Approach

We conducted a multilevel meta-analysis based on the approach outlined by Hox (2002). The units of analysis are samples of students. A random-effects model was fitted, using the MLwiN statistical software package A drawback of the random components model is that the results obtained may be less robust than outcomes obtained when applying the fixed effects model. This is especially true in the case of a relatively small number of units (samples in the present case). For illustrative purposes, we will also report outcomes based on the fixed effect model. In this way it is possible to indicate to what extent findings based on both models yield different findings.

In the two-level analyses conducted, the upper level relates to the variance between samples and the lower level relates to the variance within each sample. The inverse error variance (i.e. the squared standard error) was used for weighing at the lowest level. We constrained the variance at this level to 1. When fitting these models the variance at the upper level expresses the amount of variation in outcomes between samples.

As a first step, we fitted a zero-model for learning time at school and homework separately to analyze the average effect across samples for learning time or homework (either as an individual level variable or a variable measured at the class or school level) and the extent to which outcomes vary significantly across samples. When the significant amount of variance across samples were found, additional moderator analyses were conducted to investigate whether the variance across samples correlates with characteristics of the sample (e.g. number of students, primary or secondary education) or the characteristics of the study (e.g. design, multilevel analyses or otherwise, controlling for cognitive aptitudes or prior achievement).

We also examined whether the effect of learning time and homework differed between separate conceptualizations, including allocated time, instructional time, time on task; for homework: time, amount, frequency (see the section on moderators below).

Finally, additional analyses were conducted based on the fixed effects model. As mentioned above, an important advantage of this approach in comparison to the random effects model is the robustness of its estimates. By applying both approaches we are able to compare the findings of the most appropriate but less robust model to those of a less appropriate but more robust model. If the finding from both approaches produce similar results, this will increase the credibility of the findings (see also Cooper et al., 2006).

Moderator Analyses

Moderator analyses were conducted to examine the degree to which the relationship between learning time or homework on the one hand and student achievement on the other could be attributed to specific sample or study characteristics. Due to the low number of samples included in the meta-analysis, these moderator variables were included as covariates in the multilevel regression analysis separately (Hox, 2002).

Different facets of time variables as moderators

For learning time at school we first investigated whether the operational definition of the time variable used in the study, being categorized either as allocated learning time, instructional time or time-on-task, had a different impact on achievement. Based on previous studies we expected that the impact of instructional or time-on-task on achievement will be stronger than the effect of allocated time on achievement.

Following De Jong et al. (2000) homework variables used in the different studies were categorized in three groups: amount of homework, homework frequency and time spent on homework. The meaning of these variables at student level might not be the same as the meaning at the classroom or school level (Trautwein & Köller, 2003; Trautwein et al., 2009). Aggregated at class or school level, a positive homework effect is found when students in classes or schools that spend more time on homework outperform students in classes or schools that do not spend that much time. At individual student level the effect of homework time on achievement is positive when students spending more time on homework have better achievement gains than their peers who do not spend that much time. Homework time at class or school level is often seen as a proxy of the homework assigned, while homework time at individual level is associated with cognitive abilities and/or motivational aspects (such as e.g. prior knowledge or study habits).

In our meta-analysis we, therefore, made a distinction between homework at individual level and homework at the classroom or school level while analyzing the effects of homework (De Jong et al., 2000; Dettmers, Trautwein et al., 2009; Trautwein, Köller, Schmitz & Baumert, 2002; Trautwein & Köller, 2003; Trautwein et al., 2009). As multi-level analysis enables estimating homework effects both at individual student level and at school/class level, our analyses provide the opportunity to compare outcomes for both cases. In earlier studies that used multilevel analyses positive associations were found at school/class level, but negative associations at individual level (Gustafsson, 2013; Trautwein et al., 2002; Trautwein & Köller, 2003). At both levels the strength of the association diminishes as control variables were used in the analysis. It appeared to be difficult to "separate" homework effects from ability and motivational factors at individual student level, and, at class level, to isolate homework from associated factors of good quality teaching. Due to this the negative association found at individual level may be a spurious one.

Sample and study characteristics used as moderators

Next to different facets of time, we also used the sample and study characteristics as moderator variables, including geographical region, level of schooling (primary, secondary schools), study design, model specification, whether or not covariates at the student level are taken into account and whether or not multilevel analysis was employed. In addition, following an approach presented by Hox (2002), we used the total sample size of the studies as a moderator variable to check on publication bias. Each type of moderator will be explained briefly below.

We examined the effects of the *geographical region* in which the studies were conducted as differences in learning time and homework practices across countries may have an impact on the size of the time-achievement association. In a previous meta-analysis by Scheerens et al. (2007) studies that investigated the impact of learning time on achievement in the Netherlands produced a significant lower effect compared to studies carried out in other countries, while for homework the effect sizes found in the United States and in the Netherlands were substantially higher compared to those in other countries. In this meta-analysis, we therefore made a distinction between European countries, North American countries and other countries.

In addition, we also examined whether the time and achievement correlation was moderated by the *level of schooling*. Cooper (1989) reported that effect sizes for the association between homework and achievement were lower for studies conducted in elementary schools than for studies carried out in middle schools. The strongest effect sizes were found in studies that were conducted in high schools. In their meta-analysis on homework, Cooper et al. (2006) also found a significant positive relationship between homework and achievement at secondary level, while the effect for primary schools depended on the effect model used in the analysis (fixed versus random effects model). Therefore it might be expected that higher effects of the homework-achievement relationship are found in secondary than in primary education (see also Trautwein et al., 2009). For learning time in school the opposite might be expected as Scheerens et al. (2007) found lower effect sizes in secondary education compared to those in primary education.

The other moderator variables refer to the *model specification*, i.e. whether or not studies have accounted for covariates at the student level (SES and cognitive aptitude/prior achievement) and to the statistical techniques conducted (whether or not multilevel analysis was conducted). It seems plausible that the use of more advanced statistical techniques (such as multilevel modeling) and controlling for confounding variables will produce more accurate but lower effect estimates (see also Seidel & Shavelson, 2009).

Publication bias is a threat to the validity of meta-analyses as studies that find significant effects might have more chance to get published (Lipsey & Wilson, 2001; Sutton in Cooper, Hedges & Valentine, 2009). Hox (2002) has suggested including sample size as a moderator variable to check for publication bias. The rationale behind this recommendation is that reports of large-scale studies are likely to be published, even if they fail to show significant results. Small-scale studies may only draw attention if they come up with significant findings. Non-significant findings from small-scale studies run the highest risk of ending up in a file drawer. A negative relation between sample size and effect size must therefore be considered a strong indication of publication bias, as this indicates that relatively large effects were found in small samples.

Results

Learning Time in Schools

Table 5.3 shows the results with regard to learning time[2]. The empty model (model 0) shows the weighted average effect size for the composite measure of learning time at school over the 16 samples included. As the findings show, on average, the effect for learning time at

[2] As mentioned earlier, the unit of analysis in the quantitative meta-analysis was the independent sample. As we averaged multiple effect size estimates reported for the sample, for each sample only one effect size estimate of the relationship between time and achievement was used in the analyses.

school on student achievement is modest (F_z = .046), but significant for α < .05). Furthermore the results show that the variance across samples of learning time at school (random effect) is statistically non-significant (p = .200). Given the lack of significant variation across samples, no moderator analyses were conducted.

Table 5.3
Parameter estimates (and standard errors) of effects of (facets of) learning time in schools on student outcomes

	(0)	(1)
Intercept	.046 (.018)*	.017 (.010)
Facet of learning time in schools (RC = Allocated time)		
Instruction time		.032 (.012)*
Time on task		.093 (.071)
Variance component at between samples level	.0042	.0029
p value	.200	.099
Variance component at within sample level	1.00	1.00

Note: For each categorical variable one category was chosen as the reference category (RC)
* = significant at .05 level

In addition, we also examined the effect of three different facets of learning time in school, including allocated time, instruction time and time on task on student outcomes (see model 1). For these analyses, we used allocated time as the reference category (intercept). As shown in Table 5.3, the intercept (.017) does not deviate significantly from zero), indicating that allocated time does not have an effect on student outcomes. Furthermore, the findings show that time on task has a stronger effect than allocated time (with an effect size of .110 (.017+ .093)) but this effect is not significant. Only the effect of instructional time reaches statistical significance (Beta= .049; α < .01). Finally, the amount of variance across samples has decreased (.0029 vs. 0042), but is still not significant (p = .099). Due to this, we did not conduct additional moderator analyses.

These findings correspond with earlier research, as far as the relative magnitudes of the three facets of learning time are concerned. It should be noted that these results are based on relatively few units of analysis, so that statistical significance depends highly on the variability between the estimates.

Homework Measured at the Individual Student Level

As mentioned above, separate analyses for homework measured at the individual student level and at the classroom/school level were conducted. In this section, the results of the analysis of the effect of homework measured at the individual student level on student achievement are reported. The next section describes the results for homework effects at the class/school level.

In Table 5.4 the results with regard to homework at pupil level are reported. The empty model shows that the weighted mean effect size of homework on student outcomes is modest (.044), but significant. Furthermore, the variance across samples of the effect is also statistically significant (p < .001).

Model 1 shows the results of the analyses that focused on differential effects of facets of homework (time spent on homework, amount of homework and frequency of homework) on student outcomes. The intercept refers to the effect of time spent on homework. The results show that the different facets of homework do not have significant effects on student outcomes. Finally, the results show that the variance across samples appeared to be significant as was the case for the null model.

Table 5.4

Parameter estimates (and standard errors) of effects of (facets of) homework at student level on student outcomes

	(0)	(1)
Intercept	.044 (.022)*	.041 (.028)
Facet of homework (RC = time spent on homework)		
Amount of homework		.050 (.038)
Frequency of homework		-.016 (.038)
Variance component at between samples level	.0080	.0088
P value	< .001	< .001
Variance component at within sample level	1.00	1.00

Note: For each categorical variable one category was chosen as the reference category (RC)
* = significant at .05 level

Given the significant amount of variance across samples, moderator analyses were conducted to investigate whether differences in the findings correlate with certain sample or study characteristics. Neither level of schooling, nor statistical technique employed or model specification, did moderate the impact of homework at individual level on achievement. To control for selection bias, we also examined the moderator effect of sample size (times 10,000; centered around the grand mean). The results show no significant moderator effect of sample size. The only significant positive effect we found was the moderator effect of geographical region. In Asian samples a stronger effect of homework (at the pupil level) on student achievements was founded (with an effect size of .114 (.015 + .099)).

Table 5.5 shows the results for homework at class/school level. As the null-model shows, the weighted average effect of homework at the school/class level (as denoted by the intercept, β = .058) is small but significant. Compared to the effect of homework at individual level, the effect of homework at school/class level is somewhat stronger, however. Furthermore, the results show that the variance across samples is statistically significant (p < .001).

We also conducted subsequent analyses (model 1) that focused on the differential effects of the three facets of homework at school/class level on student achievements. The results show that we did not find significant effects of time spent on homework and amount of homework at the schools/class level. Only the effect of frequency of homework reaches statistical significance. The effect size for frequency of homework is .067 (.009 + .058).

Finally, the results show that variance across samples appeared to be significant for both models.

Table 5.5

Parameter estimates (and standard errors) of effect of (facets of) homework at class/school level on student outcomes

	(0)	(1)
Intercept	.058 (.014)*	.009 (.014)
		.065 (.046)
Facet of homework (RC = time spent on homework)		
Amount of homework		
Frequency of homework		.058 (.021)*
Variance component at between samples level	.0022	.0022
p value	< .001	< .001
Variance component at within sample level	1.00	1.00

Note: For each categorical variable one category was chosen as the reference category (RC)
* = significant effect at .05 level

Given the significant amount of variance across samples, moderator analyses were conducted to investigate whether differences in the findings correlate with certain characteristics of the samples or the study. The results of these analyses showed no significant moderator effects. Neither a significant effect was found for the for the size of the sample (times 10,000; centered around the grand mean) to check for selection bias.

In order to verify the robustness of our analyses the results obtained by testing random effect models, were compared to the outcomes of testing fixed effect models.

Although the assumptions underlying this model do not apply in the present case, an important advantage of the fixed effect model in comparison to the random effects model is the robustness of its estimates. By applying both approaches we are able to compare the findings of the most appropriate but less robust model to those of a less appropriate but more robust model. If the finding from both approaches produce similar results, this will increase the credibility of the findings. Tables 5.6 and 5.7 show the results of the comparison between the two approaches. The findings indicate that the effects of learning time and homework are positive as expected, but quite small. If a 95% confidence interval is drawn up for the estimates obtained with the random effects model, the findings based on the fixed effects model fall within these intervals. The fact that both approaches produce similar results increases the credibility of our findings.

Table 5.6
Comparison of fixed-effects model and random-effects model (estimate and standard error)

	Estimate		Standard Error	
	Fixed effects	Random effects	Fixed effects	Random effects
Learning time (n =16)	.029***	.046*	.003	.018
Homework individual (n=19)	.068***	.044*	.001	.022
Homework class level (n=12)	.053**	.058***	.003	.014

* = significant at .05 (one-tailed), **= significant at .01 (one-tailed), *** = significant at .001 (one-tailed)

Table 5.7
Comparison of fixed-effects model and random-effects model (95% confidence interval)

	95% confidence interval (Fixed effects)		95% confidence interval (Random effects)	
	Lower bound	Upper bound	Lower bound	Upper bound
Learning time (n =16)	.024	.034	.010	-.083
Homework individual (n=19)	.067	.070	.002	.089
Homework class level (n=12)	.049	.059	.030	.086

Conclusion and Discussion

The aim of this meta-analysis was to increase our knowledge on the effectiveness of time for schooling and teaching as well as homework. We examined both the general overall impact of learning time and homework as well as the differential effects of facets of learning time and homework.

Our analyses yielded small positive significant overall effects of learning time in schools, homework at individual level and homework at class/school level. The differential effects that were found for the different facets of learning time in schools and those for homework at class level were in the expected direction. A stronger effect was found for time on task than for instruction time and allocated time. At class/school level the amount and frequency of homework appeared to be more positively related to achievement than time on homework. The findings for homework time at student level were less conclusive. The analyses revealed just as many positive as negative relationships between homework time and achievement. Moreover, the effects for most of the facets of learning time, homework at pupil level and homework at class/school level did not reach statistical significance. We only found significant effects for instructional time (facet of learning time in schools) and frequency of homework (facet of homework at class level). The non-significance of the other facets should be interpreted against the background of the small n (number of effects) for each facet.

One explanation for the small effects we found might be the way of measurement and operationalization of the time and homework variables. In most of the studies included in our meta-analyses, learning time in schools or homework was measured by means of global self-reports in teacher or student questionnaires, often with one item only rather than a composite. Classroom observations, video analysis, teacher or student logs were used seldom. A combination of data sources might provide more reliable measures and deeper insight into the impact of time and homework (see Kunter & Baumert, 2006). The operationalizations used should also be considered. In the studies included in our meta-analyses, learning time in schools included a broad range of definitions (such as instruction time per week or year, number of days in the school year, percentage of pupils attending full day schools, the percentage of pupils on task every five minutes), most of them rather distal to executive learning activities. In the same vein, homework time was frequently operationalized rather narrowly as minutes spent per week.

A second line of interpretation is the methodology used in the meta-analyses. For learning time in schools, three earlier meta-analyses used comparable methodology in the sense of inclusion criteria and control for moderator variables. Two of them, Kyriakides et al. (2010) and Scheerens et al. (2007) found overall effects somewhat stronger than in the present study. Seidel & Shavelson (2007) reported a lower overall effect size. The latter researchers applied strict inclusion criteria in their meta-analysis and only included studies that had controlled for student prerequisites. For homework, just one synthesis (Scheerens et al., 2007) is available to which we could compare our findings. The overall effect found by Scheerens et al. is comparable to the findings for homework in the present meta-analysis. The authors did not distinguish between homework at individual and class/school level. However, conducting moderator analyses and thereby controlling for student prerequisites diminished the effect of homework, which even became negative.

A third line of explanation for the small effects found might be related to the lack of control for potentially important "other" variables, in studies included in earlier meta-analyses. These other variables might be student prerequisites, or other effectiveness enhancing factors, like content covered and instructional quality. When such variables are included in the analyses strong reduction in effect sizes might occur, see e.g. Boonen, Van Damme & Onghena (2013), Dettmers et al. (2009) and Trautwein (2007). Results from a study by Van Ewijk and Sleegers, (2010) further support this notion. These authors compared educational effectiveness studies in which only one focal independent variable was used (in their case the socio economic status of the students, SES) to studies where a range of variables was included next to the focal variable. The effect size for SES appeared to be significantly smaller in the second case, namely when a range of other independent variables was included. Some evidence from our own results points in the same direction. In the vote count analysis (not described in this article but published in Hendriks et al., 2014, we found that studies that included only learning time and no other effectiveness enhancing variables in the model specification, showed a sizeable higher percentage of positive significant effects than studies that also included other effectiveness enhancing variables. However, the studies on learning time in schools included in the meta-analyses do not corroborate the findings of the vote count. Only four out of twelve studies included just learning time and no other variables in the model specification. In the other eight studies a range of effectiveness enhancing variables were included. The model specification thus varied considerably across the studies and no clear conclusions could be drawn which variables matter most (see also Cooper et al., 2006 who report similar results, based on their meta-analysis of 30 correlational survey studies that applied multivariate analysis techniques on homework).

Studies that build on comprehensive models to be tested in diverse school contexts are needed to get a better understanding which factors matter most in the time or homework-achievement relationship (for example see the work of Trautwein and colleagues (Trautwein et al., 2006; Trautwein, 2007; Trautwein et al., 2009 as well as the study by Boonen, Van Damme & Onghena, 2013).

Our ambitions to improve the nuance and depth in meta-analyses of time and homework effects were hampered by the relatively small number of effect sizes, both overall and in the sub-sets defined by specific facets, as well as when conducting moderator analyses.

Moderator analyses of study and sample characteristics were conducted for homework at individual level and homework at class/school level. Moderator effects for learning time in schools could not be estimated because of the non-significant amount of variance across samples. The analyses yielded only one statistical significant effect. For homework at pupil level the analyses showed that associations with achievement were typically stronger and more positive in samples with students from Asia than in samples with students form Europe or North America. Although there are some indications that homework is differentially associated with achievement across countries, the evidence is rather scarce. The majority of homework studies so far, especially those in which multilevel analysis was applied have been conducted in a limited number of European countries. An interesting exception is a recent study on homework effects across countries conducted by Dettmers et al. (2009). This study's results are in line with ours, showing that, at school level, homework is positively associated with achievement. At pupil level the effects found were mixed, producing negative, positive, as well as non-significant effects.

A possible limitation of our meta-analyses is that we concentrated on studies which used cognitive outcomes as the dependent variable. We did not take into consideration the impact of time on other types of outcomes. While there is substantial evidence of the impact of learning time in schools and homework on student achievement, relatively little research has been reported regarding the effects on non-cognitive outcomes. In their meta-analysis Seidel and Shavelson (2007) found higher effects of learning time in schools on learning processes (regulation of learning activities) and motivational affective outcomes as compared to those on achievement, but the effects were based on just a few studies. Several authors emphasized the relevance and opportunity homework assignments offer for stimulating self-regulated learning, including meta-cognitive strategies and influences of motivation and self-efficacy (Kitsantas & Zimmerman, 2009; Trautwein et al., 2002; Winne &

Nesbit, 2010; Zimmerman & Kitsantas, 2005). In two studies these latter authors tested a path model in which students' self-efficacy for learning and perceived responsibility beliefs were included as mediating variables between homework and achievement. We see this as an interesting direction for future research.

Finally, the quality and relevance of research on time and homework effects is likely to become stronger, to the degree that the designs of the basic studies become stronger, either through a more frequent application of experimental designs, or by means of other approaches to strengthen causal inference, e.g. Diris (2012), Gustafsson (2013).

References

Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice, 18*, 5-25. doi:10.1080/0969594X.2010.513678

Boonen, T., Van Damme, J., & Onghena, P. (2013). Teacher effects on student achievement in first grade: which aspects matter most? *School Effectiveness and School Improvement, Policy and Practice*, *25*, 126-152. doi:10.1080/09243453.2013.778297

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Bushman, B. J., & Wang, M. C. (2009). Vote-counting procedures in meta-analysis. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.) (pp. 207-220). New York: Russell Sage Foundation.

Canadian Council on Learning (2009). *A systematic review of literature examining the impact of homework on academic achievement*. Retrieved from: http://www.ccl-cca.ca/pdfs /SystematicReviews/SystematicReview_HomeworkApril27-2009.pdf

Cool, V. A., & Keith, T. Z. (1991). Testing a model of school learning: Direct and indirect effects on academic achievement. *Contemporary Educational Psychology, 16*, 28-44. doi: 10.1016/0361-476X(91)90004-5

Cooper, H. (1989). *Homework*. White Plains, NY: Longman.

Cooper, H, Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987-2003. *Review of Educational Research*, *76*, 1-62. doi:10.3102/00346543076001001

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.) (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

De Jong, R., Westerhof, K. J., & Creemers, B. P. M. (2000). Homework and student math achievement in junior high schools. *Educational Research and Evaluation*, *6*, 130-157. doi: 10.1076/1380-3611(200006)6:2;1-E;F130

Dettmers, S., Trautwein, U., & Lüdtke, O. (2009). The relationship between homework time and achievement is not universal: Evidence from multilevel analyses in 40 countries. *School Effectiveness and School Improvement, 20,* 375-405. doi:10.1080 /09243450902904601

Diris, R. (2012). *The economics of the school curriculum* (doctoral dissertation). Maastricht: Universitaire Pers.

Fraser, B. J., Walberg, H. J., Welch, W. W., & Hattie, J. A. (1987). Syntheses of educational productivity research. Special Issue of *the International Journal of Educational Research*, *11*(2).

Gustafsson, J-E. (2013). Causal inference in educational effectiveness research: a comparison of three methods to investigate effects of homework on student achievement. *School Effectiveness and School Improvement*, *24*, 275-295. doi:10.1080/09243453.2013 .806334

Hendriks, M. A., Luyten, J. W., Scheerens, J., & Sleegers, P. J. C. (2014). Meta-analyses. In J. Scheerens (Ed.), *Effectiveness of time investments in education* (SpringerBriefs in Education) (pp. 55-142). Cham: Springer. doi:10.1007/978-3-319-00924-7_2

Hox, J.J. (1995). *Applied multilevel analysis* (2nd ed.). Amsterdam: TT-publikaties.

Hox, J. (2002). *Multilevel analysis techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Keith, T. Z. (1982). Time spent on homework and high school grades: A large-sample path analysis. *Journal of Educational Psychology, 74,* 248-253. doi:10.1037/0022-0663.74.2.248

Keith, T., Diamond-Hallam, C., & Fine, J. (2004). Longitudinal effects of in-school and out-of-school homework on high school grades. *School Psychology Quarterly, 19,* 187-211. doi:10.1521/scpq.19.3.187.40278

Kitsantas, A., & Zimmerman, B. J. (2009). College students' homework and academic achievement: The mediating role of self-regulatory beliefs. *Metacognition and learning, 4*, 97-110. doi:10.1007/s11409-008-9028-y

Kohn, A. (2006). Abusing research: The study of homework and other examples. *Phi Delta Kappan, 88*(1), 8-22.

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9,* 231-251. doi:10.1007/s10984-006-9015-7

Kyriakides, L., & Creemers, B. P. M. (2008). A longitudinal study on the stability over time of school and teacher effects on student outcomes. *Oxford Review of Education, 34*, 521-545. doi:10.1080/03054980701782064

Kyriakides, L., Creemers, B., Antoniou, P., & Demetriou, D. (2010). A synthesis of studies searching for school factors: implications for theory and research. *British Educational Research Journal*, *36*, 807-830. doi:10.1080/01411920903165603

Lam, J. F. (1996). *Tijd en kwaliteit in het basisonderwijs* [Time and quality in primary education]. (doctoral dissertation). Enschede: University of Twente.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Mooring, A. M. (2004). High school chemistry homework: What works? (Unpublished manuscript). Williamsburg, VA: The College of William and Mary.

Oberon (2009). *Een oriëntatie naar verlengde onderwijstijd. Inrichting en effecten*. Utrecht: Oberon.

Paschal, R. A., Weinstein, T., & Walberg, H. J. (1984). The effects of homework on learning: A quantitative synthesis. *Journal of Educational Research*, *78*, 97-104.

Peterson, R. A., & Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology, 90*, 175-181. doi:10.1037/0021-9010.90.1.175

Scheerens, J. (2014). Introduction. In J. Scheerens (Ed.), *Effectiveness of time investments in education* (SpringerBriefs in Education) (pp. 1-5). Cham: Springer.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Elsevier Science Ltd.

Scheerens, J., Luyten, H., Steen, R., & Luyten-de Thouars, Y. (2007). *Review and meta-analyses of school and teaching effectiveness*. Enschede: Department of Educational Organisation and Management, University of Twente.

Scheerens, J., Seidel, T., Witziers, B., Hendriks, M., & Doornekamp G. (2005). *Positioning and validating the supervision framework.* Enschede: University of Twente, Department of Educational Organization and Management.

Seidel, T., & Shavelson, R. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77, 454-499. doi:10.3102/0034654307310317

Stallings, J., & Mohlman, G. (1981*). School policy, leadership style, teacher change and student behavior in eight schools*. Final report to the National Institute of Education, Washington D.C.

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.) (pp. 435-452). New York: Russell Sage Foundation.

The Core Academic Learning Time Group (2002). *Review of the literature on "time and learning."* Poway, CA: Poway Unified School District and Poway Federation of Teachers.

Trautwein, U. (2007). The homework-achievement relation reconsidered: Differentiating homework time, homework frequency, and homework effort. *Learning and Instruction, 17*, 372-388. doi:10.1016/j.learninstruc.2007.02.009

Trautwein, U., & Köller, O. (2003). The relationship between homework and achievement - Still much of a mystery. *Educational Psychology Review, 15,* 115-145. doi:10.1023 /A:1023460414243

Trautwein, U., Köller, O., Schmitz, B., & Baumert, J. (2002). Do homework assignments enhance achievement? A multilevel analysis in 7th-grade mathematics. *Contemporary Educational Psychology, 27*, 26-50. doi:10.1006/ceps.2001.1084

Trautwein, U., Lüdtke, O., Schnyder, I., & Niggli, A. (2006). Predicting homework effort: Support for a domain-specific, multilevel homework model. *Journal of Educational Psychology, 98*, 438-456. doi:10.1037/0022-0663.98.2.438

Trautwein, U., Schnyder, I., Niggli, A., Neumann, M., & Lüdtke, O. (2009). Chameleon effects in homework research: The homework-achievement association depends on the measures used and the level of analysis chosen. *Contemporary Educational Psychology, 34*, 77-88. doi:10.1016/j.cedpsych.2008.09.001

Van Ewijk, R., & Sleegers, P. (2010). Peer ethnicity and achievement: A meta-analysis into the compositional effect. *School Effectiveness and School Improvement*, *21*, 237-265. doi:10.1080/09243451003612671

Walberg, H. J. (1986). Synthesis of research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (Vol. 3, pp. 214-229). New York, NY: Macmillan.

Winne, Ph. H., & Nesbit, J. C. (2010). The psychology of academic achievement. *The Annual Review of Psychology*, *61*, 653-678. doi:10.1146/annurev.psych.093008.100348

Zimmerman, B. J., & Kitsantas, A. (2005). Homework practices and academic achievement: The mediating role of self-efficacy and perceived responsibility beliefs. *Contemporary Educational Psychology, 30,* 397-417. doi:10.1016/j.cedpsych.2005.05.003

Studies Used for Meta-Analysis

Chen, S. Y., & Lu, L. (2009). After-school time use in Taiwan: effects on educational achievement and well-being. *Adolescence, 44*, 891-909.

Dettmers, S., Trautwein, U., Lüdtke, O., Kunter, M., & Baumert, J. (2010). Homework Works if Homework Quality Is High: Using Multilevel Modeling to Predict the Development of Achievement in Mathematics. *Journal of Educational Psychology, 102*, 467-482. doi: 10.1037/a0018453

Engin-Demir, C. (2009). Factors Influencing the Academic Achievement of the Turkish Urban Poor. *International Journal of Educational Development, 29*, 17-29. doi:10.1016 /j.ijedudev.2008.03.003

Eren, O., & Henderson, D.J. (2008). The impact of homework on student achievement. *Econometrics Journal, 11*, 326-348. doi:10.1111/j.1368-423X.2008.00244.x

Flowers, T. A., & Flowers, L. A. (2008). Factors affecting urban African American high school students' Achievement in reading. *Urban Education, 43*, 154-171. doi:10.1177/0042085907312351

Fuchs, Th. & Woessman, L. (2007). What accounts for international differences in student performance? A re-examination using PISA data. *Empirical Economics, 32*, 433-464. doi:10.1007/s00181-006-0087-0

House, J. D. (2005). Classroom Instruction and Science Achievement in Japan, Hong Kong, and Chinese Taipei: Results from the TIMSS 1999 Assessment. *International Journal of Instructional Media, 32*, 295. Retrieved from HighBeam Research: http://www.highbeam.com/doc/1G1-137861420.html

Hungi, N. (2008). Examining Differences in Mathematics and Reading Achievement among Grade 5 Pupils in Vietnam. *Studies in Educational Evaluation, 34*, 155-164. doi:10.1016/j.stueduc.2008.07.004

Hungi, N., & Postlethwaite, N. T. (2009). The key factors affecting grade 5 achievement in Laos: Emerging policy issues. *Educational Research for Policy and Practice, 8*, 211-230. doi:10.1007/s10671-009-9070-9

Kitsantas, A., Cheema, J., & Ware, H. W. (2011). Mathematics achievement: The role of homework and self-efficacy beliefs. *Journal of Advanced Academics, 22*, 310-339. doi:10.1177/1932202X1102200206

Kotte, D., Lietz, P., & Lopez, M. M. (2005). Factors influencing reading achievement in Germany and Spain: Evidence from PISA 2000. *International Education Journal, 6*, 113-124.

Kupermintz, H., Ennis, M. ., Hamilton, L.S., Talbert, J.E., & Snow, R.E. (1995). Enhancing the validity and usefulness of large-scale educational assessments. 1. Nels-88 mathematics achievement. *American Educational Research Journal*, *32*, 525-554. doi:10.3102 /00028312032003525

Kyriakides, L., & Creemers, B. P. M. (2008). A longitudinal study on the stability over time of school and teacher effects on student outcomes. *Oxford Review of Education, 34*, 521-545. doi:10.1080/03054980701782064

Leseman, P. P. M., Sijsling, F. F., & Vries, E. M. de (1992). Zorgbreedte en instructiekenmeken: aanknopingspunten voor de preventie van functioneel analfabetisme in het LBO. *Pedagogische Studiën*, *69*, 371-387.

Lockheed, M. E., & Komenan, A. (1989). Teaching quality and student achievement in Africa: the case of Nigeria and Swaziland. *Teaching & Teacher Education*, *5*, 93-115. doi:10.1016/0742-051X(89)90009-7

Lubbers, M. J., Van der Werf, M. P. C., Kuyper, H., & Hendriks, A. A. J. (2010). Does homework behavior mediate the relation between personality and academic performance? *Learning and Individual Differences, 20*, 203-208. doi:10.1016 /j.lindif.2010.01.005

McDonald Connor, C., Son, S-H., Hindman, A. H., & Morrison, F. J. (2005). Teacher qualifications, classroom practices, family characteristics, and preschool experience: Complex effects on first graders' vocabulary and early reading outcomes. *Journal of School Psychology, 43*, 343–375. doi:10.1016/j.jsp.2005.06.001

Muijs, D., & Reynolds, D. (2000). School effectiveness and teacher effectiveness in mathematics: Some preliminary findings from the evaluation of the mathematics enhancement programme (primary). *School Effectiveness and School Improvement, 11*, 273-303. doi:10.1076/0924-3453(200009)11:3;1-G;FT273

Natriello, G., & McDill, E. L. (1986). Performance standards, student effort on homework and academic achievement. *Sociology of Education*, *59*, 18-30.

Pugh, G., & Telhaj, S. (2003, September). *Attainment effects of school enmeshment with external communities: Community policy, church/religious influence, and TIMSS-R mathematics scores in Flemish secondary schools.* Paper presented at the European Conference on Educational Research, Hamburg.

Rossmiller, R. A. (1986. April). *School resources, home environment, and student achievement gains in grades 3-5*. Paper presented at AERA, San Francisco. Madison: Wisconsin Center for Education Research.

Taylor, B. M., Pearson, P. D., Peterson, D. S., & Rodriguez, M. C. (2003). Reading growth in high-poverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. *Elementary School Journal, 104*, 3-28.

Teodorović, J. (2011). Classroom and School Factors Related to Student Achievement: What Works for Students? *School Effectiveness and School Improvement, 22*, 215-236. doi:10.1080/09243453.2011.575650

Trautwein, U. (2007). The homework-achievement relation reconsidered: Differentiating homework time, homework frequency, and homework effort. *Learning and Instruction*, *17*, 372-388. doi:10.1016/j.learninstruc.2007.02.009

Trautwein, U., Köller, O., Schmitz, B., & Baumert, J. (2002). Do homework assignments enhance achievement? A multilevel analysis in 7th-grade mathematics. *Contemporary Educational Psychology, 27*, 26-50. doi:10.1006/ceps.2001.1084

Trautwein, U., Lüdtke, O., Schnyder, I., & Niggli, A. (2006). Predicting homework effort: Support for a domain-specific, multilevel homework model. *Journal of Educational Psychology, 98*, 438-456. doi:10.1037/0022-0663.98.2.438

Trautwein, U., Schnyder, I., Niggli, A., Neumann, M., & Lüdtke, O. (2009). Chameleon effects in homework research: The homework-achievement association depends on the measures used and the level of analysis chosen. *Contemporary Educational Psychology, 34*, 77-88. doi:10.1016/j.cedpsych.2008.09.001

Uguroglu, M., & Walberg, H. J. (1986). Predicting achievement and motivation. *Journal of Research and Development in Education*, *19*(3), 1-12.

Wagner, P., Schober, B., & Spiel, C. (2008). Time students spend working at home for school. *Learning and Instruction, 18*, 309-320. doi:10.1016/j.learninstruc.2007.03.002

Won, S. J., & Han, S. (2010). Out-of-school activities and achievement among middle school students in the U.S. and South Korea. *Journal of Advanced Academics, 21*, 628-661. doi:10.1177/1932202X1002100404

Table A1

Summary of the 12 studies (16 samples) on learning time in schools used in the meta-analysis

Authors (publication year)	Sample	Countries in sample	School type	Measure of learning time	Facet	Outcome measure	Students (N)	Statistical technique employed	Prior attainment/ ability included as a covariate	SES included as a covariate	Effect (Fisher z)
Eren & Henderson (2008)		USA	S	Weekly hours of math	A	GAA	6913	R	Yes	Yes	-.004
Fuchs & Woesmann (2007)		31 countries	S	Instruction time	A	Lang, Math, Science	96416	R	No	Yes	.031
Hungi (2008)		Vietnam	P	Per cent full day	A	Lang, Math	72376	ML	No	Yes	.030
Kotte et al. (2005)		Germany	S	Weekly lessons of instruction	A	Lang, Math	5073	ML	No	Yes	.000
		Spain	S	Weekly lessons of instruction	A	Lang, Math	6214	ML	No	Yes	.000
Kyriakides & Creemers (2008)		Cyprus	P	Actual time spent teaching	I	Math	1579	ML	Yes	Yes	.013
Lockheed & Komenan (1989)		Nigeria	S	Days in the school year	A	Math	700	R	No	No	.240
		Swaziland	S	Days in the school year	A	Math	587	R	No	No	-.025
McDonald Connor et al. (2005)		USA	P	Time spent on instruction	I	Lang	735	SEM	Yes	Yes	.060
Muijs & Reynolds (2000)	Year 1	UK	P	% Time on task	T	Math	656	ML	Yes	Yes	.020
	Year 3	UK	P	% Time on task	T	Math	709	ML	Yes	Yes	-.020
	Year 5	UK	P	% Time on task	T	Math	763	ML	Yes	Yes	.085
Pugh & Telhaj (2003)		Belgium	S	Minutes per week teaching math	I	Math	5259	R	Yes	Yes	.050
Taylor et al. (2003)		USA	P	% Time on task	T	Lang	792	ML	Yes	No	.270
Teodorovic (2011)		Serbia	P	Time spent on whole-class instruction	I	Lang, Math	4875	ML	Yes	Yes	.062

Authors (publication year)	Sample	Countries in sample	School type	Measure of learning time	Facet	Outcome measure	Students (N)	Statistical technique employed	Prior attainment/ ability included as a covariate	SES included as a covariate	Effect (Fisher z)
Uroglu & Walberg (1986)		USA	P	Time	I	Lang, Math, Science	250	R	Yes	Yes	.043

P: primary education; S: secondary education; A: Allocated time: I: Instruction time; T: Time on task; Lang: language; GAA = general academic achievement: P: Pearson correlation analysis, R: Regression analysis; ML: Multilevel analysis, SEM: Structural Equation Modeling

Table A2

Summary of the 17 studies (19 samples) on homework at individual level used in the meta-analysis

Authors (publication year)	Sample	Countries in sample	School type	Measure of homework	Facet	Outcome measure	Students (N)	Statistical technique employed	Prior attainment/ ability included as a covariate	SES included as a covariate	Effect (Fisher z)
Chen & Lu (2009)		Taiwan	S	Daily hw hours	T	GAA	10347	P	No	No	.218
Dettmers et al. (2010)		Germany	S	Time spent on math hw week	T	Math	3483	ML	Yes	Yes	-.020
Engin-Demir (2009)		Turkey	S	Hw completion	F	Grade	719	R	No	Yes	.060
Flowers & Flowers (2008)		USA	S	Hours spent hw	T	Lang	15362	R	No	Yes	.191
Hungi (2008)		Vietnam	P	Hw corrected	F	Lang; math	72376	ML	No	Yes	.055
Hungi & Postlethwaite (2009)		Laos	P	Hw given	A	Lang; math	7450	ML	No	Yes	.055
Kitsantas et al. (2011)		USA	S	Time spent math hw	T	Math	5200	ML	No	Yes	-.080
Kyriakides & Creemers (2008)		Cyprus	P	Time spent math hw	T	Math	1579	ML	Yes	Yes	.020
Lubbers et al. (2010)		Nether-lands	P	Hw time per week	T	Lang; math	9740	P	No	No	-.040
Natriello & McDill (1986)		USA	S	Time spent hw day	T	Lang	12146	PA	Yes	Yes	.127
Rossmiller (1986)		USA	P	Time spent hw day	T	Lang	95	R	No	No	-.129
Teodorovic (2011)		Serbia	P	Time spent hw	T	Lang; math	4857	ML	Yes	Yes	.072
Trautwein (2007)	Study 3	Germany	S	Time spent on assignment	T	Math	483	PA	Yes	No	-.030
Trautwein et al. (2006)		Germany	S	Time spent on assignment	T	Lang; math	414	R	Yes	No	-.087
Trautwein et al. (2009)		Germany	S	Hw time	T	Lang; math	1275	P	No	No	-.041
Wagner et al. (2008)	Study 2	Austria	S	Hw time per week	T	Lang; math	246	P	No	No	.151
	Study 3	Austria	S	Hw time per week	T	Lang; math	342	P	No	No	.110
Won & Han (2010)	Korea	Korea	S	Amount of time	A	Math	4918	R	No	Yes	.130
	USA	USA	S	Amount of time	A	Math	6772	R	No	Yes	-.020

P: primary education; S: secondary education; hw= homework; T= homework time, F = homework frequency, A = Amount of homework, Lang: language; GAA: general academic achievement; P: Pearson correlation analysis, R: Regression analysis; ML: Multilevel analysis, PA = path analysis, SEM: Structural Equation Modeling

Table A3

Summary of the 10 studies (12 samples) on homework at class level used in the meta-analysis

Authors (publication year)	Sample	Countries in sample	School type	Measure of learning time	Facet	Outcome measure	Students (N)	Statistical technique employed	Prior attainment/ ability included as a covariate	SES included as a covariate	Effect (Fisher z)
House (2005)	Japan	Japan	S	Frequency teacher gives hw	F	Science	4745	P	No	No	.038
	Hong Kong	Hong Kong	S	Frequency teacher gives hw	F	Science	5179	P	No	No	.075
	Taiwan	Taiwan	S	Frequency teacher gives hw	F	Science	5772	P	No	No	.122
Hungi (2008)		Vietnam	P	Average hw corrected	F	Lang, math	72376	ML	No	Yes	.085
Hungi & Postlethwaite (2009)		Laos	P	Average hw corrected	F	Lang, math	7450	ML	No	Yes	.030
Kupermintz et al. (1995)		USA	S	Time on hw	T	Math	5460	R	Yes	Yes	.020
Kyriakides & Creemers (2008)		Cyprus	P	Amount of hw assigned	A	Math	1579	ML	Yes	Yes	.017
Leseman et al. (1992)		Netherlands	S	Amount of hw from language	A	Lang	2605	ML	No	Yes	.141
Trautwein (2007)	Study 3	Germany	S	Hw time	T	Math	483	PA	Yes	No	-.030
Trautwein et al. (2002)		Germany	S	Frequency of hw	F	Math	1976	ML	Yes	Yes	.115
Trautwein et al. (2009)		Switzerland	S	Hw frequency	F	Math	1275	P	No	No	-.005
Wagner et al. (2008)		Austria		Average weekly working time at home for school	T	Lang, math	236	P	No	No	.169

P: primary education; S: secondary education; hw= homework; T= homework time, F = homework frequency, A = Amount of homework; Lang: language; GAA = general academic achievement; P: Pearson correlation analysis, R: Regression analysis; ML: Multilevel analysis, PA = path analysis, SEM: Structural Equation Modelling

6

Conclusions and discussion

School effectiveness research seeks to identify and investigate those malleable conditions at classroom, school and above school level that might directly or indirectly affect the learning outcomes of students. As from the early phases of school effectiveness research there was a great interest in reviews and later also meta-analyses that compile the start-of-the-art knowledge and show evidence on factors that are associated with better student achievement.

This dissertation study builds on these previous reviews and meta-analyses and is aimed to further contribute to the school effectiveness knowledge base. The dissertation consists of four reviews and meta-analyses of key effectiveness enhancing factors that operate at different levels of the school effectiveness models: school size, school leadership, evaluation and assessment and time and homework. In the reviews and meta-analyses we not only searched for the direct effects but the indirect effects that key factors might have on achievement, were examined as well. As far as the techniques of meta-analysis is concerned, different methods of analysis were applied depending on the data provided in the primary studies. As such the dissertation study not only contributes to the cumulative knowledge base but gives also insight in the many methodological and conceptual challenges in meta-analysis and school effectiveness research.

In this chapter we start with summarizing the research questions, the methods and the main findings of the four meta-analyses separately, followed by a discussion of some general conclusions concerning the magnitude of the effects found and the type of relationships (direct, indirect and nonlinear effects) searched for. Besides limitations the chapter also describes a number of suggestions for future school effectiveness research and meta-analysis.

Summary of Main Findings from Meta-Analyses and Research Reviews

School Size: Review of Direct and Indirect Effects on Cognitive, Non-Cognitive and School Organizational Outcomes

In Chapter 2 the effects of school size on a variety of student, teacher, parents’ and school organizational outcomes were investigated. The research synthesis in this chapter sought to answer the following three questions: What is the impact of school size on cognitive and non-cognitive outcomes? What is the “state of the art” of the empirical research on economies of size? What is the direct and indirect impact of school size, conditioned by other school context variables on student performance (where indirect effects are perceived as influencing through intermediate school and instruction characteristics)? As the nature of the data did not permit a meta-analysis, a vote count procedure was applied to include the most prominent indications of the directions of effects across studies and samples. The vote count was based on 84 studies, which relate to 107 samples producing 277 effect sizes.

With regard to academic outcomes the vote count revealed that “size does not matter”. For academic achievement the majority of effects reported (62%) failed to reach statistical significance. Evidence favoring smaller schools (18% of the effects) was most frequently reported. Just 9 per cent of the effects reported appeared to be significant positive (favoring larger schools), and for 11% of the effects the relationships were best described as a curvilinear effect (with in secondary education the optimal school size found between 1100 and 1400 students on average).

Concerning non-cognitive outcomes the vote count distinguished between six types of outcomes: attitudes of students and teachers towards school, participation of students, teachers or parents, school safety, student absence and dropout, other student outcomes (attitudes towards self or learning; engagement) and, school organization, teaching and learning. The overall picture that emerges from the vote count analysis on non-cognitive outcomes favours smaller schools, as nearly half of all the school size effects reported in the reviewed studies are statistically significant and negative. This evidence however is more convincing for attitudes of students and teachers towards schools and participation, then for other non-cognitive outcomes.

With respect to attitudes of students and teachers towards school 14 studies (including 14 samples) were reviewed. The evidence clearly suggests a positive impact of small schools on student and teacher attitudes, as 63% of the 24 effects were found to be statistically significant and negative. For participation the evidence was even more consistently in favour of small schools as 80% of the 13 effects found turned out to be negative and significant.

For safety and absence and drop-out the available evidence reports negative linear relationships as well, but the evidence is less convincing than for attitudes and participation. For safety 24 studies (54 effects) were reviewed and the percentage of statistically not significant effects found almost equalled the percentage of significantly negative effects (41% and 39% respectively). The remaining effects were either curvilinear or significantly positive. For absence and drop-out the same direction of result held true, with 43 per cent of the effects reported found to be non-significant, and another 43 per cent significantly negative. The remaining effects were either positive or curvilinear.

No clear evidence was found when attitudes towards self and learning or engagement were the dependent variable neither when the impact of school size on school organization, teaching and learning was estimated. Six studies were reviewed that relate to student engagement or attitudes towards self or towards learning. Most findings (44% of the effects) appeared not be statistically significant, although negative (33%) and curvilinear effects (22%) were reported as well. With respect to school organization, teaching and learning the same pattern became visible. Four studies were included in the review. Of the eighteen effects reported 61% were found not to be significant, 33% appeared to be negative and 6% (one effect) curvilinear.

With respect to the second research question addressed in this chapter the research evidence is quite clear-cut. All five studies that examined the relationship between school size and per pupil expenditure revealed a similar pattern: cost per pupil tend to decline as school size increases. This appeared to be especially the case for relatively small schools, with more modest reductions in costs for schools of average size or larger. This conclusion is based on studies that only controlled for student achievement or student graduation, as student population characteristics were not controlled for.

Only four out of the 84 studies included in the research synthesis applied indirect effect models and examined the role of intermediate conditions at school and classroom level (i.e. school climate, attendance policies, extracurricular participation and organizational learning) that mediate the relationship between school size and outcomes. Although these studies provide some support that the positive effects of school size are mediated by school and class conditions, more research is needed to get insight in the role of preconditions and a consistent set of intermediate conditions that account for the assumed relationship between school size and cognitive and non-cognitive outcomes.

School Leadership Effects Revisited; A Review of Empirical Studies Guided by Indirect-Effect Models

Previous meta-analyses of school leadership usually focussed on the direct impact of leadership on student outcomes. School effectiveness studies that assume school leadership behaviour affects achievement indirect through intervening school organisational and instructional conditions took some time to get off the ground, with the results of one of the first empirical studies (the leadership for organizational learning and student outcomes (LOSLO) project (Mulford, 2003; Silins & Mulford, 2004) being published in the early 2000s.

As the indirect effect models could provide us with more insight into the paths and mechanisms through which school leadership practices may impact on student achievement, they were the focus of the review study in Chapter 3. The following research questions were addressed: What is the total (direct and indirect) effect of school leadership on student achievement? What are the most promising paths and intermediate variables in indirect effect models that study the impact of school leadership on student achievement? To answer these questions, we analysed 15 studies conducted between 2005 and 2010 that addressed indirect-effect models of school leadership. In all these studies, structural equation modelling was used to investigate the direct and indirect effects of leadership on achievement. A quantitative meta-analysis was applied as well as a narrative review, providing information on the intermediary school and instructional variables that might play a role in explaining indirect school leadership effects.

The non-weighted summary effect of the 34 effects found in the 15 studies was modest (r = .031), and statistically not significant. The weighted summary effect was r = .048. However, when one publication with outlying negative effects was excluded, the mean effect across the remaining 14 studies became r = .06, and statistically significant. Across the 14 studies the weighted mean effect was almost equal to the non-weighted mean: r = .061.

To answer the question related to the most promising paths and intermediate variables we first calculated the effect for each direct or indirect path between a leadership measure and an achievement measure in the indirect effect models. The indirect effects were calculated as the product of the association of leadership with a particular intermediate variable and the association of the intermediate variable(s) and student outcomes. In total 36 direct and indirect paths were distinguished, with effects that vary from r = -.32 to r = .25. Remarkable outcomes are the rather high negative direct effects of r = -.27 and r = -.32 in two studies (De Maeyer, Rymenans, Van Petegem, Van den Bergh & Rijlaarsdam, 2007; Ten Bruggencate, 2009 respectively). These negative associations sometimes are interpreted as compensatory action of schools and school leaders as a reaction to low student performance. However, given the correlational nature of the studies in question, these interpretations are rather speculative. On the other hand, the most positive promising paths and intermediate variables in the indirect effect models referred to academic climate (r = .25) , school conditions (r = .24) and instructional practices(r = .16, r = .14).

Further examination of the intervening variables showed a wide range of conditions being selected in the 15 studies included in the review. The intermediate variables that "seemed to work" could be broadly organized into organizational capacity (improvement focus, standard setting, quality of student support, professional capacity of the staff, student monitoring and feedback), teachers' commitment and cooperation, academic climate and instructional conditions. Quite a few studies addressed a broad spectrum of effectiveness enhancing school factors, while in other studies the intermediate variables selected were targeted on specific school conditions. Intermediate variables covering instructional practices were lacking in most of the indirect effect models. Just three studies included aspects of teaching effectiveness. The conceptual models applied in these studies build on key assumptions of integrated educational effectiveness models, in which conditions at school level are seen as a means to support and facilitate conditions at classroom level. The results of the study by Heck & Moriyama (2010) provided support for the causal ordering of school leadership, instructional variables and student outcomes.

The review showed that further quantitative and qualitative work would be needed to strengthen the knowledge base on indirect leadership effect models and obtain more detailed information how the respective intermediary conditions work (and possibly interact) in influencing student achievement. As far as intermediary variables is concerned, more specific connection to instructional effectiveness in school leadership effect studies seems to be a promising direction for further research.

Effects of Evaluation and Assessment on Student Outcomes: A Review and Meta-Analysis

From the early days of school effectiveness research in reviews and meta-analyses evaluation and assessment has been mentioned as one of the 'core' correlates of effective school and instructional conditions, and this has not changed until today. What is more, evaluation and assessment are increasingly considered as potential levers of change that could assist with decision-making and continuous improvement at student, class, school and above school level. The review and meta-analysis that is presented in Chapter 4 focused on the impact of evaluation and assessment as effectiveness enhancing conditions at school and classroom level. The meta-analysis included 7 studies on evaluation at school level, 14 studies on evaluation at class level and 6 studies that examined the impact of assessment. The outcome variable in all studies was student achievement.

In the meta-analysis a random effects model was employed, following the procedures provided by Lipsey and Wilson (2001). A vote count procedure was applied as well to permit the inclusion of studies that did not provide sufficient information to calculate an effect size.

The meta-analyses yielded statistically significant but small positive effects for evaluation at school and evaluation at class level (r = .070 and r = .073 respectively[1]), while the average effect size for assessment was almost zero and non-significant (r = .01). Results of the vote count were in the same direction. Across the three variables (evaluation at school level, evaluation at class level and assessment), the vote counts indicated a weak general predominance of positive effects compared to negative effects (28 % versus 4%), while a substantially higher percentage of positive effects was found for evaluation at school level (46% versus 1%).

The Q test of the homogeneity of the effect sizes was conducted to examine whether there was significant variability across studies. The Q test proved to be statistically significant for evaluation at class level. However, as the number of effect sizes included in the sample was quite small (n = 15) it was decided not to conduct further moderator analysis. For evaluation at school level and assessment the Q statistic appeared to be statistically non-significant. The reason for this might be the low number of effects in the samples (n = 7 and n = 7 respectively).

As evaluation and assessment have a place in rational planning models the concept of the evaluative cycle was used to analyse the conceptualizations and operationalizations of evaluation of the predictor variables used in the primary studies. Five phases were distinguished: 1) setting the objectives and standards of the evaluation, 2) data collection, 3) evaluative interpretation of the data, 4) feedback and 5) use, implementation, and action. The results of the conceptual analysis showed that a thorough and complete application of the evaluative cycle was rarely addressed in any of the studies included in the review. Data collection and use, implementation and action were the phases most commonly addressed (in studies conducted at both class and school level), as well as the phase of feedback for evaluation (class level only). Hardly any empirical research was found on the processes by which teachers and school leaders noticed and interpreted data. This finding is touched upon also by other authors (see e.g. Bennett, 2011) who suggests that interpreting or making inferences is only just beginning to become integrated into definitions of formative assessment.

The quality of the evaluative cycle and its impact on teaching and learning rests in part on the attitudes, knowledge and skills that teachers and school leaders have in evaluation and assessment and the strategies they use. Creating effective evaluation and assessment cycles at all educational levels requires capacity building and professional learning at both teacher, school and above school level. Further research therefore is recommended to understand what types of teacher collaboration and professional development opportunities will enhance effective evaluation and assessment practices.

[1] Fisher's Z comparable to the correlation coefficient r for small effects (r<.25)

Time for Schooling and Teaching: A Meta-Analysis on the Effects of Learning Time in Schools and Homework

Time for schooling and teaching is considered one of the key policy amenable variables to improve educational outcomes and the quality of teaching and learning. Time for teaching and learning are at the core of the educational effectiveness models and has meaning at all educational levels distinguished. Previous meta-analyses yielded small to medium positive effects on the impact of learning time in schools and homework.

The study presented in Chapter 5 attempted to validate findings on the effectiveness of learning time in schools and homework available up to 2005 with the findings from recent studies. In the meta-analysis we considered both the general overall effect of time as well as the differential effects of facets of learning time and homework. In the case of learning time in schools a distinction was made between allocated time, instruction time and time on task. Three types of measures of homework were distinguished: homework frequency, homework time and amount of homework. Moreover, when analysing the effects of homework we also distinguished between homework at the individual student level and homework at the school/class level, as the meaning of homework at these two levels might not be the same (see e.g. Trautwein & Köller, 2003).

The meta-analysis included 12 studies (16 samples) on learning time in schools, 17 studies (19 samples) for homework at individual level and 10 studies (12 samples) for homework at class/school level. A multilevel meta-analysis was applied based on the approach outlined by Hox (2002). A random effects model was applied. In case of significant variance across samples, moderator analyses were conducted to examine the degree to which the association between learning time or homework and student achievement could be attributed to certain characteristics of the samples or the study. Due to the small number of samples in the meta-analyses, the moderator variables were included as covariates in the regression analysis separately.

The meta-analyses yielded small positive significant overall effects of learning time in schools and homework at both individual and class/school level (r = .046, r = .044 and r = .058 respectively) as well as for two of the nine facets of time (i.e. instruction time: r = .048 and homework frequency at school/class level: r = .067). The non-significance of the other facets should be seen against the background of the relatively small number of effects in the sub-sets for each facet. Statistical significance then highly depends on the variability between the estimates.

The overall effects found both for time for learning and homework, are lower than those reported in most previous meta-analyses. The differential effects the analysis yielded for the different facets of learning time in schools and those for homework at class level were in the expected direction. A stronger effect was found for time on task than for instruction time and allocated time. At class/school level the amount and frequency of homework appeared to be more positively related to achievement than time on homework. The findings for homework time at student level were less conclusive. In our meta-analysis a small positive effect was found. What is more, in the vote count (see the publication

Effectiveness of Time Investments in Education, Scheerens, 2014a), the analyses revealed just as many positive as negative relationships between homework time and achievement. This is in contrast to earlier European studies in which homework was examined at more than one level (see e.g. Trautwein, Schnyder, Niggli, Neuman & Lüdtke, 2009) and multilevel analysis produced negative or negligible effects at individual student level. It is in line with the study by Dettmers, Trautwein and Lüdtke (2009), in which homework effects were examined across 40 countries. In this study at pupil level, the effects found were mixed as well, yielding negative, positive and non-significant effects.

Moderator analyses of study and sample characteristics were conducted for homework at individual level and homework at class/school level. Moderator effects for learning time in schools could not be estimated because of the non-significant amount of variance across samples. The analyses yielded only one statistical significant effect. For homework at pupil level the analyses showed that a stronger and more positive effect of homework on student achievement was found in samples with students from Asia than in samples with students from Europe or North America.

A possible limitation of this meta-analysis is that we did not take into consideration the impact of learning time and homework on other student outcomes than achievement. The meta-analysis by Seidel and Shavelson (2007) revealed higher effects of learning time in schools on learning processes (regulation of learning activities) and motivational affective outcomes than those on student achievement. Zimmerman & Kitsantas (2005) examined the role of homework on student's self-regulation and responsibility for their learning. These variables were found promising mediating factors between homework and achievement.

General Conclusions and Discussion

In this section, some general conclusions are drawn on the basis of the proceeding findings. In addition an attempt is made to reflect on the magnitude and direction of effects found in this dissertation study, as well as on the type and direction of the relationships that were examined in the primary studies and meta-analyses. In this section limitations and implications for further research are discussed as well.

Interpretation of Effect Sizes

In general the effects found in this dissertation study could be considered negligible to small, both in comparison to what Cohen (1998) classifies a small effect size[2] as well as compared to the results of some other recent meta-analyses. For those factors for which we were able to conduct meta-analyses (school leadership, evaluation and time), small positive, statistically significant effects on achievement were found for six of the seven effectiveness enhancing variables included in review (see Table 6.1). The review did not indicate a significant effect of assessment on student learning.

[2] According to Cohen (1998) small effects are in the order of r = .10, medium effects r = .30 and large effects r = .50 or higher.

Table 6.1

Mean effects from meta-analyses on school leadership, evaluation and time (this dissertation)

	Average effect (r^3)	Number of studies/samples included	Number of effects included
School leadership			
Leadership	.061*	14	34
Evaluation			
Evaluation at school level	.073**	7	7
Evaluation at class level	.073***	15	15
Assessment	.005	7	7
Time			
Learning time in schools	.046*	16	31
Homework at class/school level	.058***	12	19
Homework at individual student level	.044*	19	30

* = significant at .05, **= significant at .01, *** = significant at .001

For those effectiveness enhancing conditions for which we applied vote counting (school size and evaluation and assessment), the impact appeared to be weak as well. For evaluation and assessment the vote count indicated a small positive effect with 28 per cent of the estimates found to be positively and statistically significant related to student learning and 4 per cent negatively related. For school size the pattern of the vote count varied with the type of outcome measure used in the studies. Effects of smaller schools appeared to be strongest for two types of non-cognitive outcomes, i.e. attitudes of students and teachers towards school (social cohesion) and participation of students or parents, while for safety and attendance the effects found tend to favor smaller schools as well. School size did not seem to matter for academic achievement. Two third of the associations between size and achievement appeared to be non-significant, and for the remaining one third of effects the number of significant negative, positive and curvilinear relationships did not differ that much from each other.

Although the effects found are small and may appear to be discouraging, this does not mean that they are neither unimportant nor unrealistic. As was suggested in Chapter 3 for leadership it might be unreasonable to expect large effects given the long causal chain between leadership actions and student achievement and the research designs that were used in most of the studies analyzed. Also, the small and insignificant estimates might be due to the limited variance in educational outcomes and school and instructional processes identified in national school effectiveness studies. The latter might be one of the reasons for the small effects that we found for (certain facets of) learning time in schools, such as allocated learning time, as there might not be much variance in a country. Moreover,

[3] Fisher's Z comparable to the correlation coefficient r for small effects (r<.25)

allocated time alone cannot produce effective learning and is only effective if time is spent on covering the right content taught and good quality teaching. Small school and instructional effects thus could be considered of significance as they might be cumulative.

Several authors argue that Cohen’s guidelines are to be considered as too conservative in the education context or even argue that there is no universal statistical guideline for judging the “educational significance” of a standardized effect size estimate (see e.g. Bloom, Hill, Black & Lipsey, 2008; Durlak, 2009; Lipsey, Puzzio, Yun, Herbert, Steinka-Fry, Cole, Roberts, Anthon & Busick, 2012). Instead effect sizes should be interpreted with respect to empirical benchmarks that are relevant to practical or substantive considerations, such as comparison with a typical learning gain during one school year in a certain subject matter outcome measure for a given target population of students. Another approach is to compare the effect sizes with the effects observed in similar types of studies (or in our case meta-analyses). A meaningful magnitude then depends on the degree to which the same type of intervention or effectiveness enhancing factor, target population, research design or outcome measures is being considered in the studies or meta-analyses that are to be used as benchmark. This is not easy as different researchers apply different inclusion and quality criteria in their meta-analyses, use different methods to construct the effectiveness enhancing variables and the dependent variables and employ different methodological techniques to calculate mean effect sizes. This becomes also visible when we compare the effect sizes found in this dissertation study to those reported in other recent meta-analyses (see Table 6.2). A first result that emerges from Table 6.2 is the variation in average effects that the various meta-analyses yielded. The second observation that can be made refers to the relatively very small effects that this dissertation study yielded. Hattie and Timperley (2007), Hattie (2009), Kyriakides, Creemers, Antoniou and Demetriou (2010) and Kyriakides, Christoforou and Charalambous (2013) all generally found substantially larger effects than those that derived from our meta-analyses. Seidel and Shavelson (2007) were the only authors that reported smaller effects. The latter authors only included studies that had controlled for student prerequisites. With the exception of learning time in schools the effects that Scheerens, Luyten, Steen and Luyten-de Thouars (2007) found are similar to those in this dissertation study.

Table 6.2

Comparison of findings from this dissertation study to recent meta-analyses on school leadership, evaluation and time

	This study	Hattie & Timperley (2007)	Scheerens et al. (2007)	Seidel & Shavelson (2007)	Hattie (2009)	Kyriakides et al. (2010)	Kyriakides et al. (2013)
School leadership							
Leadership	.06		.06		.18	.07	
Evaluation							
Evaluation at school level	.07		.06			.13[1]	
Evaluation at class level	.07	.35[2]		.01			
Assessment	.01			.02	.17	.18[3]	.34[4]
Time							
Learning time in schools	.05		.15	.03[5]		.16[6]	.35[7]
Homework at class/school level	.06		.07				
Homework at individual student level	.04						

Variable heading of effectiveness enhancing variable in meta-analysis in case of divergent conceptualization: [1]Evaluation of school policy on teaching and actions taken for improving teaching practice, [2]Feedback, [3]Student assessment (school level), [4]Student assessment (class level), [5]Learning time, opportunity to learn and homework, [6]School policy on the quantity of teaching, [7] Time management

Possible Explanations for the Rather Weak Effects Found in the Meta-Analyses

As already mentioned above, the rather weak effects that this dissertation study revealed might be inherent to the nature of school effectiveness research and its predominance on correlational research design. Most of the school effectiveness research is naturalistic in nature and investigates the natural variance in ‘real life’ schools and classrooms. From the perspective of practical significance this can be seen as an advantage because of the high ecological validity in this type of studies. From the perspective of internal validity however correlational studies are more vulnerable which might result in lower effect sizes as compared to studies that apply a quasi-experimental or experimental research design. In (quasi-)experimental studies, researchers reduce the complexity and concentrate on specific, better controlled interventions, in which it is possible to have rigorous control for confounding variables. When interventions in (quasi-)experimental studies are implemented with high fidelity these type of studies are more likely to demonstrate teaching effects of greater magnitude. (Quasi-)experimental studies however are weaker in terms of their ecological validity and the often higher effects found should be interpreted from this perspective as well.

In our meta-analyses a correlational research design was applied in most of the studies sampled. For school leadership and time all the studies included in the meta-analyses were correlational. For evaluation the majority of studies had a correlational research design as well, and only four out of the 21 studies included applied an experimental or (quasi-) experimental research design, which each yielded relatively high effects (varying between r = .191 and r = . 236, see Chapter 4).

In other meta-analyses (Kyriakides et al., 2010; Kyriakides et al., 2013; Scheerens et al., 2007; Seidel & Shavelson, 2007) a mix of correlational, experimental and quasi-experimental studies was sampled as well. These authors also examined the moderating effect of research design. While Kyriakides et al. (2010) and Kyriakides et al. (2013) indeed found a positive and significant moderating effect of experimental studies for four of the five factors for which the moderator analysis was possible, this moderating effect was less in evidence in the meta-analysis by Scheerens et al. (2007). Hattie is less explicit about the types of studies sampled. However, a predominance of intervention studies included in his meta-analyses might be expected as Hattie reported that “most of the successful effects come from innovations and these effects of innovations might not be the same as the effects of teachers in regular classrooms” (Hattie, 2009, p. 6).

The relatively low effects derived from correlational studies could be explained by their focus on more distal aspects of schooling and teaching relative to the execution of learning activities (see Bolhuis, 2003) and the widely different operationalizations thereof by means of global self-reports in teacher or student questionnaires. Also, the value-added models used in correlational studies mostly are assessing status and measure change in student achievement less frequently (see also Rowan, Correnti & Miller, 2002). In longitudinal models the effects might be more substantial as they could be cumulative and unfold over time.

A second line of explanation for the small effects found can be related to the model specification and the different types of value added multivariate models applied in the primary studies included in the meta-analyses. In school effectiveness research the complexity of the relationship between an effectiveness enhancing factor and outcomes is modelled by including other potential relevant variables that earlier research findings or theory suggest to have an effect on student achievement. These variables might be covariates at student level, compositional variables or other effectiveness enhancing conditions. Many student, teacher, class, school and above school-level variables may play a role. When such factors are included in the analyses, strong reduction in effect sizes might occur or the effects the factor of interest might even disappear, see e.g. Boonen, Van Damme and Onghena (2013) and Garrett, Newman, Elbourne, Bradley, Noden, Taylor and West (2004).

In the vote count analysis of studies on learning time in schools (not in this dissertation study but included in Scheerens, 2014a), we indeed found a sizeable higher percentage of positive significant effects for studies that included only learning time and no other effectiveness enhancing variables in the model specification compared to studies that also included other effectiveness enhancing variables. We unfortunately were not able to corroborate this finding in the meta-analysis. Due to the variation in model specification across studies and the limited number of studies included, no clear conclusions could be drawn which variables matter most in the learning time-achievement relationship. The potential importance of model specification however is also illustrated in other studies (see e.g. Cooper, 2006; Garrett et al., 2004). Within the context of school size effects this latter author refers to the studies of Bickel and Howley (2000) and Bickel, Howley, Williams and Glascock (2001). Using the same data from more than 300 schools, these authors found that the inclusion of a greater number of exploratory variables and the application of hierarchical modelling in the second study resulted in a non-significant association between school size and student achievement, whereas a statistically significant positive relationship was found in the first study.

It is therefore important that future effectiveness studies build on earlier research, substantively and methodologically. But, although comprehensive models and established theory of educational effectiveness are available, most studies seem to have a more explorative rather than a confirmative character. A review of 109 studies (Nordenbo et al., 2009 in Scheerens, 2013) showed that only a minority of the studies was anchored in theory or based on more elaborated conceptual models such as those by Creemers (1994) and Creemers and Kyriakides (2008). What is more, the lack of congruence in the way the key factors are conceptualized and operationalized and the application and measurement thereof, is not conducive for the development of a robust knowledge base. The latter became apparent from the analysis of primary studies in the meta-analyses in this dissertation study as well (see in especially the Chapters 4 and 5), and might also be a cause for the low effects and the many insignificant findings.

A third reason why we might not have expected strong and consistent effects lies in the loose coupling between the hierarchical levels in educational organisations. Loose coupling (Weick, 1976 in Scheerens, 2014b) seems to be contradictory to the concept of school effectiveness that depends on the rational idea of optimal attainment of educational goals (often student achievement).

According to Scheerens (2014b) the school as loosely coupled organization might be seen as an explanation for the ineffectiveness of schools rather than for the effectiveness. It might also explain the limited malleability of some of the effectiveness enhancing factors and the relatively low effects found in meta-analyses for these factors. While in school effectiveness models school level conditions are assumed to be facilitating and buffering conditions of effective classroom conditions, in a loosely coupled system, with a relatively small interdependence between subunits, teachers have considerable autonomy and there is less urgency for coordination and leadership (Scheerens, Glas & Thomas, 2003).

In Chapter 3 the idea of loosely coupling was elaborated upon to explain the rather small and indirect effects that are found in school leadership effectiveness research. In this chapter it was discussed that schools have many substitutes for educational leadership and that under normal circumstances a lean kind of control of teachers and teaching would be sufficient.

The Impact of Effectiveness Enhancing Factors: Direct, Indirect and Non-Linear Effects

In this dissertation study we included studies that examined the direct impact of effectiveness enhancing conditions as well as studies that searched for the indirect effects. We must note that in the majority of studies included in our reviews and meta-analyses a regression or multilevel model was applied searching for direct and linear effects of an effectiveness enhancing condition on pupil achievement. In these studies, various teacher, class, school, and above school-level variables are added to the model and controls for student background and composition are made. Although multilevel models are more adequate to depict the modelling of school effectiveness than regression as they take into consideration the hierarchical nature of schooling these models do not account for the indirect effects that the school might have on student achievement through mediating factors.

Therefore, besides including studies that examined a direct linear relationship we also searched for studies investigating the impact of indirect effects, in especially those studies that examined the effects of leadership and school size. Given the long causal chain between the independent variable and achievement for these two variables, an indirect effects model, in which the impact of school size or leadership on achievement is modelled as mediated by intermediate school and/or instruction characteristics, is more valid. In school leadership studies, research methods like structural equation modelling that enable researchers to explicate and test the causal ordering of factors have been more applied frequently than in studies that examine the direct and indirect effects of school size.

For school leadership we were able to analyse 15 indirect effect school leadership studies. Although the findings of our review did not indicate a more substantial effect of leadership on achievement than those in some earlier meta-analyses of direct effect studies, the review gained insight in the intermediary conditions that might play a role in explaining leadership effects. Especially a more specific connection to instructional practices in school leader effectiveness studies seems to be a promising direction for further research. The findings of promising indirect paths to leadership effects in the review matched key assumptions of integrated effectiveness models where conditions at school level are seen as relevant to the extent that they support and facilitate instructional conditions at classroom level. As such, school leadership studies, testing these direct and indirect effects, are to be seen as interesting examples and promising direction to the further development of conceptual models and theory in school leadership effectiveness research (see also Heck & Moriyama, 2010; Scheerens, 2012).

As far as school size is concerned, the evidence on mediation models we could gather was much more limited. Only four studies in the review on school size addressed indirect effect models and provided some information on the role of potential mediating variables. For school size more research that tries to explain the role of potential intermediary conditions is thus badly needed. However, compared to school leadership, studying school size effectiveness is more complex as far as the choice of dependent variables is concerned, and some of these might also function as relevant intermediating variables in a longer causal chain. E.g. participation and social cohesion might be selected as non-cognitive outcomes but could also be modelled as mediating variables that either facilitate or impede the impact of school size on student achievement.

Moreover, in the research on the impact of school size other forms of relationships are reported as well, especially nonlinear relationships. In these studies a curvilinear relationship is assumed and researchers searched for the optimal school size. The studies sometimes revealed an "inverted U" relationship, which implies that to be most effective schools should be neither too small nor too large. In some other cases the studies failed to demonstrate a nonlinear relationship. This does not necessarily mean that a nonlinear relationship does not exist as the non-significance of the findings might be due to difficulties of establishing enough variation in either the dependent variable or the range of school enrolment included in the study. Although some reviewers report an optimal school size for some of the dependent variables (see Luyten, Hendriks & Scheerens, 2014), in our review we were not able to provide optimal school sizes. The ranges of size that studies included differed too much and optimal sizes were mainly reported for studies modelling achievement as the dependent variable. Instead we took into consideration all the different types of modelling the relationship between size and the cognitive and non-cognitive outcomes and reported the findings for each type of modelling separately.

Although in this dissertation study nonlinear relationships were taken into consideration for school size effects only, they might exist for the impact of other effectiveness factors as well. Creemers, Kyriakides and Sammons (2010) e.g. refer to classroom climate and teacher management, but the amount of learning time in schools and homework might operate in a nonlinear fashion as well (see also Cooper, Robinson & Patall, 2006).

Limitations

Meta-analysis is a well-established method of summarizing research evidence from a range of independent studies that address a related research question and has some advantage over traditional methods such as narrative review and vote count analysis (Borenstein, Hedges, Higgins & Rothstein, 2009). Compared to traditional review methods one of the most distinguished features of a meta-analysis is the conversion of the results of an individual study in an effect size statistic. By standardizing effect sizes of individual studies researchers are able to compare across different studies as well as to integrate the results and to establish an average effect size (see also Chapter 1 of this thesis). On the other hand the design and implementation of meta-analysis requires stricter conditions than traditional review methods. Such restrictions resulted in some limitations that will be discussed below.

Publication bias is a threat to the validity of meta-analysis and narrative reviews as studies with statistically significant effects are more likely to be published and published studies are more likely to be included in systematic reviews (Borenstein et al., 2009). In the reviews and meta-analyses we have tried to account for this by applying thorough search procedures. But although our search procedures resulted in many studies for possible review just a small number of potential studies met the inclusion criteria for the meta-analyses (see Table 6.1 for an overview of the number of studies included the meta-analyses on school leadership, evaluation and time).

The possible indication of bias however was not confirmed when we checked for publication bias in the meta-analysis on time. Following a procedure recommended by Hox (2002), we included sample size as a covariate in the moderator analysis. The results of these analyses showed no significant moderator effect of sample size to check for selection bias.

A major reason that we had to drop a substantial number of studies from the meta-analyses was that authors of the primary studies failed to report effect sizes or did not provide the summary statistics needed to compute an effect size statistic. Studies were also excluded because of selective reporting. In the latter case researchers only reported the significant effects and did not show the effects that were non-statistically significant.

These issues were of influence on the methods of analyses that we applied and the conclusions that can be drawn. The first choice we made was to employ vote counting as well as vote counting permits inclusion of those studies that reported on the significance and direction of the association between the effectiveness enhancing factor and an outcome measure but that do not provide sufficient information to permit the calculation of an effect size. The vote counting method is the simplest and most conservative method for combining results of independent studies and comes down to counting the numbers of positive significant effects, negative significant effects and non-significant effects (all based on a two-sided test with $p<.05$). The vote count procedure has some disadvantages as it is less powerful for combining effect sizes and sample sizes. Also, when a study consists of more effect sizes these are counted each separately in the vote-counting. In that case the study with more effects has a larger impact on the results than a study in which only one effect is reported. Vote counting therefore should not be seen as a full blown alternative to the quantitative synthesis but rather as a “next best” solution which we choose to apply given the limitations of the studies sampled.

As far as meta-analysis is concerned we were able to apply different models in the studies (i.e. a ‘traditional’ random effects mode following the procedures provided by Lipsey and Wilson (2001) and multilevel meta-analysis following procedures by Hox (2002), see respectively Chapter 4 and 5). The limited number of studies however, hampered our ambitions to improve the depth and nuance of earlier meta-analyses on evaluation and time and interpreting the mean effect sizes found in this dissertation study therefore needs some caution. The results of the moderator analyses, as far as it was possible to conduct these, should be considered with caution as well. The number of effects included in each category was relatively small and due to the low number of samples included in the meta-analysis the moderator variables were included as covariates in the multilevel regression analysis separately.

Implications for Further Study

Meta-analysis is assumed to be a powerful tool for appraising the cumulative knowledge base in a field. As this dissertation showed it is not easy to explain the divergence in mean effect sizes in various meta-analyses when the impact of the same effectiveness factor is examined. In order to generate a more valid and credible knowledge base, researchers could learn from the meta-analyses presented in this dissertation and other past meta-analyses and use it as a starting point for follow up meta-analyses and primary studies. In this light replication studies, especially replications of those meta-analyses that yielded relatively large effects, could provide insight into the methodological choices and judgments that are made in the original meta-analyses. (e.g. problem definition, choice of selection and inclusion criteria, coding procedure, research design, calculation of effect sizes, weighing of effect sizes, data analysis methods and reporting results).

In addition we recommend meta-analysts to explicitly document the decisions taken and the procedures adopted in all stages of the meta-analysis and to make these available to the scientific community. The need to thoroughly describe methods and results also applies to researchers of primary studies. Although several authors suggested guidelines about the kind of information that researchers of primary and meta-analysts should typically provide (see e.g. Ahn, Ames & Myers, 2012; Creemers, Kyriakides & Sammons, 2010; Harwell & Maeda, 2008) these should become more generally available and also included in academic courses on research methodology. When these guidelines are taken into consideration, this will benefit the quality of future studies and meta-analyses and the inferences that can be drawn from these reviews.

References

Ahn, S., Ames, A. J. & Myers, N. D. (2012). A review of meta-analyses in education: methodological strengths and weaknesses. *Review of Educational Research*, *82*, 436-476. doi:10.3102/0034654312458162

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*, 5-25. doi:10.1080/0969594X.2010.513678

Bickel R., & Howley, C. (2000). The influence of scale on school performance: a multi-level extension of the Matthew principle. *Education Policy Analysis Archives, 8*. Retrieved from: http://epaa.asu.edu/epaa/v8n22/

Bickel, R., Howley, C., Williams, T., & Glascock, C. (2001). High school size, achievement equity, and cost: Robust interaction effects and tentative results. *Education Policy Analysis Archives*, *9*(40). Retrieved from http://epaa.asu.edu/ojs/article/view/369/495

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*, 289-328

Bolhuis, S. (2003). Towards process-oriented teaching for self-directed lifelong learning: A multidimensional perspective. *Learning and Instruction, 13*(3), 327-347.

Boonen, T., Van Damme, J., & Onghena, P. (2013). Teacher effects on student achievement in first grade: which aspects matter most? *School Effectiveness and School Improvement, Policy and Practice*, *25*, 126-152. doi:10.1080/09243453.2013.778297

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: Wiley.

Cohen, J. (1998). *Statistical power analysis for the behavioral sciences*. 2nd edition. Hillsdale, NJ: Lawrence Erlbaum.

Cooper, H, Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987-2003. *Review of Educational Research*, *76*, 1-62. doi:10.3102/00346543076001001

Creemers, B. P. M. (1994). *The effective classroom.* London: Cassell.

Creemers, B. P. M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness.* London and New York: Routledge.

Creemers, B. P. M., Kyriakides, L., & Sammons, P. (2010). *Methodological advances in school effectiveness research.* London: Routledge.

De Maeyer, S., Rymenans, R., Petegem, P. van, Bergh, H. van den, & Rijlaarsdam, G. (2007). Educational leadership and pupil achievement: The choice of a valid conceptual model to test effects in school effectiveness research. *School effectiveness and School Improvement, 18*, 125-145. doi:1080/09243450600853415

Dettmers, S., Trautwein, U., & Lüdtke, O. (2009). The relationship between homework time and achievement is not universal: Evidence from multilevel analyses in 40 countries. *School Effectiveness and School Improvement, 20,* 375-405. doi:10.1080 /09243450902904601

Durlak, J. A. (2009). How to select, interpret and calculate effect sizes. *Journal of Pediatric Psychology, 34*(9), 917-928. doi:10.1093/jpepsy/jsp004

Garrett, Z., Newman, M., Elbourne, D., Bradley, S., Noden, P., Taylor, J., & West, A. (2004). Secondary school size: A systematic review. In *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Harwell, M., & Maeda, Y. (2008). Deficiencies of reporting in meta-analyses and some remedies. *The Journal of Experimental Education*, *76*, 403-428.

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

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112. doi:10.3102/003465430298487

Heck, R. H., & Moriyama, K. (2010). Examining relationships among elementary schools' contexts, leadership, instructional practices, and added-year outcomes: A regression discontinuity approach. *School Effectiveness and School Improvement, 21*, 377-408. doi:10.1080/09243453.2010.500097

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Lawrence Erlbaum Associates.

Kyriakides, L., Creemers, B., Antoniou, P., & Demetriou, D. (2010). A synthesis of studies searching for school factors: implications for theory and research. *British Educational Research Journal*, *36*, 807-830. doi:10.1080/01411920903165603

Kyriakides, L., Christoforou, C., & Charalambous, C. L. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education*, *36*, 143-152. doi:10.1016/j.tate.2013.07.010

Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Lipsey, M. W ., Puzzio, K., Yun, C., Herbert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). Translating the statistical representation of the effects of education interventions into more readily interpretable forms. (NCSER 2013-3000). Washington DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved from: http://ies.ed.gov/ncser/

Luyten, H., Hendriks, M. A., & Scheerens, J. (Eds.) (2014). *School size effects revisited* (SpringerBriefs in Education). Cham: Springer.

Mulford, B. (2003). *School leaders: Changing roles and impact on teacher and school effectiveness.* Paper commissioned by the Education and Training Policy Division for the activity Attracting, Developing and Retaining Effective Teachers. Paris: OECD. Retrieved from: http://www.oecd.org/dataoecd/61/61/2635399.pdf

Nordenbo, S. E., Holm, A., Elstad, E., Scheerens, J., Soegaard Larsen, M., Uljens, M., . . . Hauge, T. E. (2009). *Research mapping of input, process and learning in primary and lower secondary schools*. Copenhagen, Denmark: Danish Clearing House for Educational Research, DPU, Aarhus University.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record, 104*(8), 1525–1567.

Scheerens, J. (Ed.) (2012). *School leadership effects revisited. Review and meta-analysis of empirical studies* (Springer Briefs in Education). Dordrecht: Springer.

Scheerens, J. (2013). The use of theory in school effectiveness research revisited. *School Effectiveness and School Improvement, 24*, 1-38. doi:10.1080/09243453.2012 .691100

Scheerens, J. (Ed.) (2014a). *Effectiveness of time investments in education* (SpringerBriefs in Education). Cham: Springer.

Scheerens, J. (2014b). Theories on educational effectiveness and ineffectiveness. *School Effectiveness and School Improvement. (*accepted)

Scheerens, J., Glas, C. A. W. & Thomas, S. M . (2003). *Educational evaluation, assessment, and monitoring. A systematic approach.* Lisse: Swets & Zeitlinger.

Scheerens, J., Luyten, H., Steen R., & Luyten-de Thouars, Y. (2007). *Review and meta-analyses of school and teaching effectiveness*. Enschede: University of Twente.

Seidel, T., & Shavelson, R. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. R*eview of Educational Research*, *77*, 454-499. doi:10.3102/0034654307310317

Silins, H., & Mulford, B. (2004). Schools as learning organisations. Effects on teacher leadership and student outcomes. *School Effectiveness and School Improvement*, *15,* 443-466. doi:10.1080/09243450512331383272

Ten Bruggencate, G. C. (2009). *Maken schoolleiders het verschil?* [Do school leaders make a difference?]. Enschede, The Netherlands: University of Twente.

Trautwein, U., & Köller, O. (2003). The relationship between homework and achievement - Still much of a mystery. *Educational Psychology Review, 15,* 115-145. doi:10.1023 /A:1023460414243

Trautwein, U., Schnyder, I., Niggli, A., Neumann, M., & Lüdtke, O. (2009). Chameleon effects in homework research: The homework-achievement association depends on the measures used and the level of analysis chosen. *Contemporary Educational Psychology, 34*, 77-88. doi:10.1016/j.cedpsych.2008.09.001

Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly, 21*, 1-19.

Zimmerman, B. J., & Kitsantas, A. (2005). Homework practices and academic achievement: The mediating role of self-efficacy and perceived responsibility beliefs. *Contemporary Educational Psychology, 30,* 397-417. doi:10.1016/j.cedpsych.2005.05.003

S

Summary in Dutch (Samenvatting)

De aanleiding tot dit proefschrift komt voort uit het vroegere onderzoeksprogramma “Effectiveness of school and training organizations” van de afdeling Onderwijsorganisatie en –management van de faculteit Gedragswetenschappen van de Universiteit Twente. De leidende onderzoeksvragen in dit programma waren tweeledig:

- Welke kenmerken van scholen en arbeidsorganisaties zijn indicatief voor hoge productiviteit en effectiviteit van onderwijs- en arbeidsorganisaties?
- Welke modellen en theorieën kunnen de werking van deze kenmerken verklaren?

Een van de onderzoekslijnen sinds de start van het onderzoeksprogramma in 1989 kan gekenmerkt worden als onderzoek dat betrekking heeft op de "grondslagen van onderwijseffectiviteitsonderzoek". Deze onderzoekslijn is erop gericht een kennisbasis te genereren van de meest relevante beïnvloedbare condities, alsook het periodiek in kaart brengen van de beschikbare empirische onderbouwing. Overzichtsstudies (reviews) en meta-analyses zijn daarbij de belangrijkste methoden die toegepast zijn. Belangrijke onderzoekpublicaties op dit gebied zijn Scheerens en Bosker (1997), Scheerens, Seidel, Witziers, Hendriks en Doornekamp (2005), Scheerens, Luyten, Steen en Luyten-de Thouars (2007) en Witziers, Bosker en Krüger (2003).

Deze dissertatie bouwt voort op deze eerdere reviews en meta-analyses en is bedoeld om verder bij te dragen aan de kennisbasis van het schooleffectiviteitsonderzoek. De dissertatie bestaat uit vier reviews en meta-analyses van effectiviteitsbevorderende kernvariabelen die opereren op verschillende niveaus van schooleffectiviteitsmodellen: schoolgrootte, schoolleiderschap, evaluatie en assessment en leertijd op school en huiswerk. In de reviews en meta-analyses is niet alleen gekeken naar de directe associaties tussen school- en instructiekenmerken en uitkomsten maar is ook specifiek aandacht geweest voor indirecte en niet-lineaire verbanden. Wat betreft de technieken voor meta-analyse zijn verschillende analysemethoden toegepast afhankelijk van de data die beschikbaar waren in de primaire onderzoeken. In die hoedanigheid draagt dit proefschrift niet alleen bij aan de cumulatieve kennisbasis wat betreft effectiviteitsbevorderende school- en instructie-kenmerken maar geeft het ook inzicht in de vele methodologische en conceptuele uitdagingen in meta-analyses en schooleffectiviteitsonderzoek.

Het proefschrift bestaat uit vier reviews en meta-analyses die afzonderlijk gelezen kunnen worden (zie de hoofdstukken 2, 3, 4 en 5), en waarvan doel, methode en resultaten hieronder kort beschreven zullen worden. De dissertatie sluit af met een conclusie en discussie waarin de balans opgemaakt wordt wat betreft de grootte en de richting van de gevonden effecten en waarbij aanbevelingen voor vervolgonderzoek gedaan worden.

Schoolgrootte

In het onderzoek naar de effecten van schoolgrootte zijn twee hoofdoriëntaties te onderkennen. Enerzijds is er de vraag naar de relatie tussen schoolgrootte en de opbrengsten van het onderwijs; we beschouwen dit als het effectiviteitsperspectief. Anderzijds wordt er gekeken naar de kosten en kosteneffectiviteit van schoolgrootte, dit is het efficiency perspectief. In de overzichtsstudie, waarvan we in hoofdstuk 2 verslag doen, ligt het accent op het effectiviteitperspectief, maar is ook het efficiency perspectief in ogenschouw genomen. Daarnaast hebben we ons gericht op een derde perspectief, dat beschouwd kan worden als een nadere uitwerking van het effectiviteitsperspectief en waarbij we getracht hebben meer inzicht te krijgen in de indirecte effecten van schoolgrootte. Dit betreft de wijze waarop de mogelijke effecten van schoolgrootte gemedieerd worden door schoolorganisatie- en instructiefactoren.

De review in hoofdstuk 2 actualiseert de beschikbare kennis over de effecten van schoolgrootte op drie soorten uitkomsten: leerprestaties, niet-cognitieve uitkomsten en de kosten per leerling. De niet-cognitieve uitkomsten die in beschouwing zijn genomen, hebben betrekking op attitudes van leerlingen en leerkrachten t.a.v. hun school (sociale cohesie), participatie van leerlingen, leerkrachten en ouders, veiligheid op school, absentie, spijbelen en voortijdig schoolverlaten, overige leerlingkenmerken zoals zelfvertrouwen en betrokkenheid, en schoolorganisatie en onderwijskwaliteit.

Het overzicht is gebaseerd op 84 onderzoekspublicaties, 107 steekproeven (samples) en 277 effecten. In de review bleek het niet mogelijk een kwantitatieve meta-analyse uit te voeren. Een reden daarvoor was dat slechts een beperkt deel van de publicaties voldoende statistische informatie bevatte om een effectgrootte te kunnen berekenen. Daarnaast bleek in veel onderzoeken de relatie tussen schoolgrootte en de uitkomstmaat vaak niet als een lineaire relatie gemodelleerd. In plaats daarvan is een log-lineaire of kwadratische relatie verondersteld, of worden verschillende categorieën van schoolgrootte vergeleken, waarvan het aantal en de verdeling van de groottes over de categorieën varieert per onderzoek. Daarom is tot een ‘vote count’ analyse besloten. Een vote count analyse bestaat in essentie uit het tellen van de aantallen positief significante en negatief significante effecten. In deze overzichtsstudie hebben we echter ook de effecten betrokken die een bepaald optimum impliceren, alsook de niet-significante effecten.

De resultaten van de vote count laten zien dat de effecten van schoolgrootte op zowel cognitieve uitkomsten (leerprestaties) als op niet-cognitieve uitkomsten (zeer) beperkt zijn. Met betrekking tot leerprestaties blijkt de meerderheid van de gerapporteerde effecten (62%) niet-significant. Slechts 18% van de effecten is significant negatief (betere resultaten in kleine scholen) en 9% significant positief (betere resultaten in grotere scholen). Voor 11% van de effecten kan de relatie het best getypeerd worden als een curvi-lineair effect met voor secundair onderwijs de optimale schoolgrootte gevonden tussen gemiddeld 1100 en 1400 leerlingen.

Wat betreft de associaties tussen schoolgrootte en niet-cognitieve uitkomsten, toont de vote count een voorkeur voor kleinere scholen, met de helft van de in de onderzoekspublicaties gerapporteerde effecten statistisch significant en negatief. De bewijslast is het meest overtuigend voor sociale cohesie (attitudes van leerlingen en leerkrachten t.a.v. hun school) en participatie van leerlingen, leerkrachten en ouders (respectievelijk 63% van de 24 effecten en 80% van de 13 effecten negatief significant). Voor veiligheid, absentie en drop-out blijkt de beschikbare bewijslast minder sterk ten gunste van kleine scholen en is het percentage statistisch niet-significante effecten ongeveer gelijk aan het percentage statistisch significant negatieve effecten (elk ongeveer 40%). Wanneer overige leerlingkenmerken (zoals zelfvertrouwen en betrokkenheid) of schoolorganisatie en onderwijskwaliteit de uitkomstenmaten zijn blijkt er geen duidelijk effect te zijn.

De uitkomsten met betrekking tot de relatie tussen schoolgrootte en de kosten per leerling wijzen op lagere kosten voor grote scholen, zij het dat het aantal onderzoeken beperkt is (n = 5), en er in de onderzoeken alleen gecontroleerd is voor leerprestaties of slaagpercentages (en dus niet voor kenmerken van de leerlingenpopulatie). Kostenbesparingen zijn het grootst indien (zeer) kleine scholen groeien of worden samengevoegd. De kosten dalen veel minder sterk als scholen van gemiddelde grootte verder in omvang toenemen.

Met betrekking tot het derde perspectief moet worden vastgesteld dat slechts in vier van de 84 onderzoekspublicaties de vraag naar indirecte effecten van schoolgrootte onderzocht is. Mediërende variabelen op school- en klasniveau die gemodelleerd zijn, zijn schoolklimaat, maatregelen om verzuim te beperken, het extra-curriculaire leren en organisatieleren. De weinige uitkomsten lijken enige empirische ondersteuning te bieden voor de gedachte dat schoolgrootte een effect op leerprestaties kan uitoefenen via een minder gunstig schoolklimaat.

De belangrijkste conclusie uit de onderzoeksliteratuur is dat de effecten die men mag verwachten van veranderingen in schoolgrootte op leerprestaties en niet-cognitieve uitkomsten bescheiden zijn en een grote variatie in onderzoeksuitkomsten laat zien. Daarnaast lijkt het effect van schoolgrootte ook afhankelijk van de nationale context. Gebleken is dat veel van de negatieve effecten van schoolgrootte vooral in Amerikaans onderzoek zijn gerapporteerd. Nader onderzoek naar schoolgrootte zou zich daarom niet alleen moeten richten op de associatie met uitkomsten, maar vooral ook op de identificatie van de context- en mediërende school- en instructievariabelen die het indirecte effect van schoolgrootte kunnen verklaren.

Schoolleiderschap

Eerdere meta-analyses naar de effecten van schoolleiderschap zijn veelal gebaseerd op onderzoeken waarin een direct verband tussen leiderschap en leerprestaties onderzocht is. Onderzoeken waarin indirecte effecten van schoolleiderschap gemodelleerd worden waren tot vrij recent nogal schaars (zie o.a. De Mayer & Rymenans, 2004). Hoofdstuk 3 betrof een meta-analyse en review naar de directe en indirecte effecten van schoolleiderschap op leerprestaties. Vijftien onderzoeken waarin de indirecte effecten van leiderschap onderzocht zijn, zijn geanalyseerd. De onderzoeksvragen die daarbij centraal hebben gestaan zijn: Wat is het totale (directe en indirecte) effect van schoolleiderschap op leerprestaties? Wat zijn de meest veelbelovende paden en intermediaire variabelen in de indirecte effectmodellen die de invloed van schoolleiderschap op leerprestaties onderzoeken?

Het gewogen gemiddelde totale effect van schoolleiderschap op leerprestaties bleek bescheiden (correlatie, r = .06) en in overeenstemming met de gemiddelde effecten die ook in meta-analyses naar direct effect onderzoeken gevonden worden. Om de vraag naar de meest veelbelovende paden en intermediërende variabelen te kunnen beantwoorden berekenden we eerst het effect voor elk direct en indirect pad tussen een leiderschaps-variabele en leerprestaties onderscheiden in de modellen. De indirecte effecten werden berekend als het product van de associatie tussen leiderschap en een intermediaire variabele en de associatie tussen de intermediaire variabele en leerprestaties. In totaal zijn 36 directe en indirecte paden onderscheiden.

Opmerkelijke uitkomsten zijn de, relatief, grote negatieve directe en indirecte effecten gevonden in twee van de 15 onderzoeken (De Mayer, Rymenans, Van Petegem, Van den Bergh & Rijlaarsdam, 2007; Ten Bruggencate, 2009). Deze negatieve associaties worden soms geïnterpreteerd als compenserende actie van schoolleiders op lage leerprestaties. In dat geval is sprake van omgekeerde causaliteit, waarbij de hoogte van de leerprestaties de intermediaire variabelen en het leiderschapsgedrag bepaalt in plaats van het omgekeerde.

De meest veelbelovende indirecte paden van schoolleiderschap naar leerprestaties blijken te verlopen via de intermediaire variabelen prestatiegericht klimaat (De Mayer et al., 2007), school condities (Leithwood & Jantzi, 2008) en instructiecondities (Heck & Moriyama, 2010). Nadere beschouwing van de intermediaire variabelen leverde een grote diversiteit aan intermediërende variabelen opgenomen in de modellen in de vijftien onderzoeken. De intermediaire variabelen die “ertoe lijken te doen” kunnen geordend worden in vier categorieën, namelijk het organisatorisch potentieel van scholen (`school capacity´), betrokkenheid en samenwerking van de leerkrachten, een prestatiegericht klimaat en met condities die te maken hebben met effectieve instructie. Opvallend is dat de laatste categorie, kenmerken van instructie, pas in de meest recente onderzoeken een plaats gevonden heeft. De conceptuele modellen die in deze laatste onderzoeken toegepast zijn duiden op een betere verbinding tussen het schoolleidersonderzoek en multilevel modellen van onderwijseffectiviteit. De resultaten van het onderzoek van Heck en Moriyama (2010) in het bijzonder vertonen empirische steun voor de causale ordening van schoolleiderschap, instructiekenmerken en leerprestaties. Verder kwantitatief en kwalitatief onderzoek is echter nodig om echt conclusies te kunnen trekken over het relatieve belang van elk van de categorieën intermediërende variabelen in indirect effect modellen van schoolleiderschap.

Evaluatie

In hoofdstuk 4 zijn de effecten van evaluatie en assessment onderzocht. Hiertoe is zowel een vote count als een meta-analyse uitgevoerd. De meta-analyse omvatte zeven onderzoeken van evaluatie op schoolniveau, 14 onderzoeken van evaluatie op klasniveau en zes onderzoeken waarin het effect van assessment is nagegaan. Leerprestaties zijn de uitkomstmaat in alle onderzoeken.

In de meta-analyse is een random effects model gebruikt uitgaande van de procedures van Lipsey en Wilson (2001). De resultaten duiden op significante maar kleine effecten van evaluatie op schoolniveau en evaluatie op klasniveau (r=.07 en r=.073 respectievelijk), terwijl het gemiddelde effect voor assessment niet-significant is en bijna nihil (r=.01). De resultaten van de vote count wijzen in dezelfde richting. Voor alle drie de variabelen duiden de resultaten van de vote count op een zwak overwicht van positieve effecten ten opzichte van negatieve effecten (28% versus 4%), terwijl voor evaluatie op schoolniveau een substantieel hoger percentage positieve effecten is gevonden (46% versus 1%). De Q test voor de homogeniteit van de effecten is gebruikt om na te gaan of er significante heterogeniteit bestaat tussen de onderzoeken. Dit bleek het geval te zijn voor evaluatie op klasniveau. Voor evaluatie op schoolniveau en assessment bleek de Q test niet significant, wat veroorzaakt kan worden door het beperkte aantal onderzoeken in de steekproeven. Gegeven echter ook het kleine aantal onderzoeken voor evaluatie op klasniveau is besloten geen verdere moderator analyses uit te voeren.

Daar evaluatie en assessment een plaats hebben in rationele planningsmodellen hebben we het concept van de evaluatieve cyclus als uitgangspunt genomen om de conceptualisering en operationalisering van evaluatie en assessment in de onderzochte onderzoeken nader te analyseren. Vijf fasen zijn onderscheiden: het vaststellen van doelen en standaarden, gegevensverzameling, het analyseren en interpreteren van de vorderingen, van leerlingen, feedback en het nemen van beslissingen.

De resultaten van de conceptuele analyse laten zien dat een diepgaande en complete toepassing van de evaluatieve cyclus nauwelijks onderzocht is in de geanalyseerde onderzoeken. De focus ligt vooral op de fasen gegevensverzameling en het nemen van beslissingen (dit betreft onderzoeken waarin zowel evaluatie op school- en klasniveau onderzocht is) als de fase van feedback (evaluatie op klasniveau). De onderzoeken bieden nauwelijks enig empirisch bewijs voor de processen waarbij leraren en schoolleiders de data analyseren en interpreteren. Dit is in overeenstemming met bevindingen van andere onderzoekers (zie bijvoorbeeld Bennet, 2011) die suggereren dat het interpreteren en het trekken van conclusies nog nauwelijks onderdeel vormt van de operationalisering van het concept van formatieve evaluatie.

Leertijd

Leertijd is een van de door beleid te manipuleren kernvariabelen om de leerprestaties en de kwaliteit van onderwijzen en leren te verbeteren. Leertijd ligt aan de basis van multilevel modellen voor instructie- en onderwijseffectiviteit. Eerdere meta-analyses tonen kleine tot gemiddelde effecten voor leertijd op school en huiswerk.

De meta-analyse in hoofdstuk 5 had tot doel de resultaten van eerdere meta-analyses naar de effecten van leertijd op school en huiswerk te valideren met bevindingen uit onderzoeken die zijn uitgevoerd tussen 2005 en 2010. In de meta-analyse hebben we zowel gekeken naar de 'overall' effecten van tijd als naar de differentiële effecten van subcategorieën van leertijd op school en huiswerk. Wat betreft leertijd op school is gekozen voor een driedeling: officiële leertijd (allocated time), instructietijd (dat wil zeggen het deel van een les dat daadwerkelijk aan instructie besteed wordt) en taakgerichte leertijd (time on task). Voor de bestudering van huiswerk op school is een onderscheid gemaakt tussen de hoeveelheid huiswerk, de tijd die aan huiswerk besteed wordt en de frequentie waarmee huiswerk gegeven wordt. Tevens werd een onderscheid gemaakt tussen onderzoeken waarin huiswerk op individueel leerlingniveau geanalyseerd is en onderzoeken waar de analyse van huiswerk op school-/klasniveau uitgevoerd is, daar de betekenis van huiswerk op deze twee niveaus conceptueel niet hetzelfde is (zie bijv. Trautwein & Köller, 2003).

De meta-analyses zijn gebaseerd op 12 onderzoeken (16 steekproeven) naar leertijd op school, 17 onderzoeken (19 steekproeven) naar huiswerk op leerlingniveau en 10 onderzoeken (12 steekproeven) naar huiswerk op school-/klasniveau. Een multilevel meta-analyse is uitgevoerd op basis van de benadering van Hox (2002), waarbij een random effects model is gebruikt en een moderator analyse is uitgevoerd.

De resultaten van de meta-analyses tonen kleine positieve en significante effecten voor de ‘overall’ effecten van leertijd op school en huiswerk op individueel en school-/klasniveau (r = .046, r = .044 en r = .058 respectievelijk), alsook voor twee van de negen subcategorieën (instructietijd: r = .046 en frequentie waarmee huiswerk gegeven wordt op school-/klasniveau: r = .067). Het feit dat de overige subcategorieën van leertijd op school en huiswerk op individueel en school-/klasniveau geen significant effect laten zien kan mogelijk verklaard worden door het relatief beperkte aantal effecten per subcategorie van tijd of huiswerk. De statistische significantie van een subcategorie hangt dan sterk af van de spreiding tussen de effecten.

De ‘overall’ effecten van leertijd op school en huiswerk zijn echter wel kleiner dan die gerapporteerd in eerdere meta-analyses. De differentiële effecten voor leertijd op school en huiswerk op school-/klasniveau zijn in de verwachte richting, met een sterker effect voor taakgerichte leertijd dan voor instructietijd en officiële leertijd. Voor huiswerk op individueel niveau is dat niet het geval. In eerder onderzoek (zie bijv. Trautwein, Schnyder, Niggli, Neuman & Lüdtke, 2009) werden daar juist negatieve effecten gerapporteerd.

Voor huiswerk op individueel niveau en huiswerk op school-/klasniveau zijn moderator analyses uitgevoerd, waarbij, vanwege het beperkte aantal steekproeven in de meta-analyses, de moderator variabelen steeds als afzonderlijk covariaat in de regressie-analyse meegenomen zijn. De moderator analyses lieten een statistisch significant effect zien. Voor huiswerk op individueel niveau tonen de analyses een iets sterker en meer positief effect als de steekproeven afkomstig zijn uit Azië dan wanneer de steekproeven uit Europa of Amerika afkomstig zijn.

Algemene conclusie en discussie

In hoofdstuk 6 zijn de bevindingen samengevat en bediscussieerd. In zijn algemeenheid kan geconstateerd worden dat de effecten die in de meta-analyses in deze dissertatie gevonden zijn beschouwd kunnen worden als verwaarloosbaar tot klein, zowel in vergelijking tot wat

Cohen (1998)[1] classificeert als een kleine effectgrootte alsook wanneer we de resultaten van de meta-analyses in deze dissertatie vergelijken met die in eerdere meta-analyses (zie respectievelijk tabel 6.1 en tabel 6.2 in hoofdstuk 6).

Alhoewel de effecten die we gevonden hebben klein zijn, en wellicht ook teleurstellend, betekent dit niet dat ze er niet toe doen of niet realistisch zijn. Voor leiderschap bijvoorbeeld is in hoofdstuk 3 gesuggereerd dat gezien de lange causale keten tussen schoolleiderschap en leerprestaties en het cross-sectionele design van de meeste onderzoeken naar schoolleiderschap geen grote effecten verwacht mogen worden. De kleine en vaak ook niet significante effecten kunnen ook een gevolg zijn van de beperkte variantie in leerprestaties en in school- en instructiekenmerken in school effectiviteits-onderzoeken die binnen de context van een land worden uitgevoerd. Tot slot kunnen kleine effecten van school en instructiekenmerken toch ook als belangrijk gezien worden omdat ze cumulatief zijn.

Daarnaast zijn meerdere auteurs van mening dat de standaard van Cohen te conservatief is voor de onderwijscontext of dat er zelfs geen universele statistische richtlijnen zijn om de statistische significantie van een effectgrootte te kunnen beoordelen (zie bijvoorbeeld Bloom, Hill, Black & Lipsey, 2008; Durlak, 2009; Lipsey, Puzzio, Yun, Herbert, Steinka-Fry, Cole, Roberts, Anthon & Busick, 2012). In plaats daarvan zouden effectgroottes beter geïnterpreteerd kunnen worden in relatie tot empirische benchmarks die relevant zijn vanuit praktische of inhoudelijke overwegingen, zoals de typische leerwinst die een bepaalde doelgroep van leerlingen gedurende een jaar onderwijs in een bepaald vak behaalt. Een andere manier is om de resultaten van de effectgroottes te vergelijken met de effecten die in soortgelijke onderzoeken (in ons geval meta-analyses) gevonden zijn. Een betekenisvolle effectgrootte is dan afhankelijk van de mate waarin bijvoorbeeld het zelfde type interventie of school- of instructievariabele, doelgroep, onderzoeksdesign, of uitkomstmaat in beschouwing genomen wordt in het onderzoek of de meta-analyse waarmee vergeleken wordt. We hebben daarom onze resultaten vergeleken met die van eerdere meta-analyses. Hattie en Timperley (2007), Hattie (2009), Kyriakides, Creemers, Antoniou en Demetriou (2010) en Kyriakides, Christoforou en Charalambous (2013) vinden allemaal grotere gemiddelde effecten dan de effecten die gerapporteerd worden in deze dissertatie. Seidel en Shavelson (2007) rapporteren kleinere effecten en met uitzondering van het effect voor leertijd op school zijn de effecten die Scheerens, Luyten, Steen en Luyten-de Thouars (2007) vinden vergelijkbaar met de effecten in deze dissertatie.

De interpretatie vervolgens is echter niet eenvoudig omdat verschillende onderzoekers verschillende inclusie- en kwaliteitscriteria hanteren in hun meta-analyses, verschillende manieren gebruiken om de manipuleerbare school- en instructiekenmerken en de uitkomstmaten te construeren en verschillende analysetechnieken gebruiken om de gemiddelde effectgroottes te berekenen.

[1] Volgens Cohen (1998) worden correlaties van .10 als klein opgevat, correlaties van .50 als middel-groot en correlaties van .80 en meer als groot.

De kleine effecten die we hebben gevonden kunnen inherent zijn aan de aard van schooleffectiviteitsonderzoek en de predominantie van een correlationeel onderzoeksdesign. Schooleffectiviteitsonderzoek is vaak naturalistisch van aard, cross-sectioneel en gebaseerd op informatie verkregen met behulp van zelfpercepties van leerkrachten en leerlingen in vragenlijsten. Een voordeel is de relatief hoge ecologische validiteit maar de interne validiteit van correlationeel onderzoek is meer kwetsbaar en gevoelig voor kleine effecten vergeleken met onderzoeken die een (quasi-)experimenteel design toepassen.

Een tweede lijn van verklaringen heeft te maken met de modelspecificatie in de primaire onderzoeken in de meta-analyses. In schooleffectiviteitsonderzoek wordt de complexiteit van de associatie tussen een school- of instructiekenmerk en de uitkomsten vaak gemodelleerd door andere potentieel relevante variabelen toe te voegen aan het onderzoeksmodel. Deze variabelen kunnen covariaten zijn op leerlingniveau, compositievariabelen en andere school- en instructiekenmerken en verschillen vaak van onderzoek tot onderzoek. Wanneer deze variabelen toegevoegd worden aan het onderzoekmodel kan dit leiden tot een sterke reductie van of zelfs het verdwijnen van het effect van school- of instructiekenmerk waar de interesse naar uitgaat (zie bijvoorbeeld Boonen, Van Damme & Onghena, 2013; Garret, Newman, Elbourne, Bradley, Noden, Taylor & West, 2004). Daarnaast maakt de grote variatie in modelspecificatie tussen de verschillende onderzoeken het niet eenvoudig om heldere conclusies te kunnen trekken welke variabelen er nu het meest toe doen. Verder blijkt er weinig consistentie in de wijze waarop de kernvariabelen geconceptualiseerd en geoperationaliseerd zijn en dit belemmert de ontwikkeling van een robuuste kennisbasis. Het lijkt derhalve raadzaam dat toekomstig schooleffectiviteitsonderzoek meer voortbouwt op eerder onderzoek, zowel conceptueel als methodologisch.

Tot slot zijn in hoofdstuk 6 ook een aantal aanbevelingen voor toekomstige meta-analyses gedaan. Gezien de grote verschillen in effecten die gevonden worden in verschillende meta-analyses zijn replicaties en in het bijzonder replicaties van meta-analyses waarin grote effecten gevonden aan te bevelen. Deze replicaties kunnen ons meer inzicht geven in de methodologische keuzes en beoordelingen die gemaakt zijn in de oorspronkelijke meta-analyses (bijvoorbeeld met betrekking tot probleemdefinitie, keuze van selectie en inclusiecriteria, codeerprocedure, onderzoeksdesign, berekening van effecten, het wegen van effecten, analysemethoden en methoden van rapportage).

Daarnaast bevelen we onderzoekers die een meta-analyse uitvoeren aan om de beslissingen die genomen worden en de procedures die gevolgd worden in de verschillende fases van de meta-analyse expliciet te documenteren en te rapporteren. Deze noodzaak om de methoden en resultaten nauwgezet te beschrijven geldt ook voor de onderzoekers van de primaire onderzoeken. Alhoewel er richtlijnen bestaan voor het soort informatie dat onderzoekers van primaire onderzoeken en meta-analyses zouden moeten opleveren (zie bijvoorbeeld Ahn, Ames & Myers, 2012; Creemers, Kyriakides & Sammons, 2010; Harwell & Maeda, 2008) is het aanbevolen deze meer algemeen beschikbaar te maken. Dit kan de kwaliteit van toekomstige meta-analyses ten goede komen, alsook de conclusies die eruit getrokken kunnen worden.

A

Acknowledgements
(Dankwoord)

Toen ik in 1995 na mijn studie Onderwijskunde bij de vakgroep Onderwijsorganisatie en -management ging werken om daar samen met collega's het ZEBO instrument (zelfevaluatie-instrument voor het basisonderwijs) te ontwikkelen was een van de eerste activiteiten een conceptuele analyse van de belangrijkste variabelen van school- en instructie-effectiviteit. Vele jaren en internationale projecten later heeft dit een voortzetting gekregen in wat uiteindelijk geresulteerd heeft in het tot stand komen van dit proefschrift.

In de allereerste plaats ben ik daarvoor veel dank verschuldigd aan mijn promotor Jaap Scheerens. Jaap, ik ben je zeer dankbaar voor de mogelijkheid die je mij geboden hebt om alsnog te kunnen promoveren, voor het vertrouwen dat je in mij getoond hebt en de ruimte die je me gegeven hebt om het proefschrift op mijn manier af te ronden. Ik dank je voor je grote betrokkenheid bij dit onderzoek en de steeds waardevolle en tijdige feedback.

Peter Sleegers, ik wil jou, als vakgroepvoorzitter, danken voor de tijd en support die je mij gegeven hebt om dit proefschrift tot een goed einde te brengen. Met jouw humor, relativering en enthousiasme wist je mij steeds weer te motiveren en inspireren, ook als het soms even lastig was. Daarnaast ben ik je zeer dankbaar voor je inhoudelijke betrokkenheid en ondersteuning bij de meta-analyse in hoofdstuk 5. Ik heb veel geleerd van jouw kritische reflecties en constructief commentaar.

Rien Steen en Hans Luyten wil ik bedanken voor hun betrokkenheid bij de analyses en de methodologie. Ook voor vragen en feedback kon ik altijd bij jullie terecht. Veel dank daarnaast voor de vele uren die we besteed hebben aan het doorspreken van de afzonderlijke onderzoeken, om vaak tot de conclusie te komen dat alweer een onderzoek niet aan de inclusiecriteria voldeed.

Mijn (oud-)collega's van de vakgroep Onderwijskunde, in het bijzonder Peter, Joseph, Ruth, Nelleke, Maaike en Carola wil ik bedanken voor hun belangstelling, betrokkenheid en stimulerende samenwerking. Het is heel inspirerend om in zo een collegiaal team te mogen werken aan de verdere uitbouw van ons HRD onderwijs en onderzoek. Carola, jou ben ik bijzonder veel dank verschuldigd voor de lay-out van dit proefschrift en het regelen van allerlei zaken rondom de promotie. Ik ben er trots op dat jij en Maaike mijn paranimfen willen zijn. Veel dank daarvoor.

Tot slot mijn 'thuis': Anne en Marlot en ook Koen. Ik ben zo trots op jullie. De gedrevenheid en volharding waarmee jullie ieder op je eigen manier je dromen en ambities proberen waar te maken is heel bijzonder om mee te maken en geeft me energie. Ik ga graag mee met de volgende stappen die jullie gaan zetten. André, dat ik jou wil bedanken is evident. En dat het hoog tijd is om weer eens vaker zelf te koken ook ☺. Ik kijk ernaar uit om weer meer tijd samen te hebben.