Accounting for previous performance of students by means of growth curves analyses to estimate the size, stability and consistency of school effects

Timmermans, Antje; van der Werf, Greetje

Published in:
Educational Research and Evaluation

DOI:
10.1080/13803611.2017.1455300

Publication date:
2018

Routledge
Taylor & Francis Group

Educational Research and Evaluation

An International Journal on Theory and Practice

ISSN: 1380-3611 (Print) 1744-4187 (Online) Journal homepage: http://www.tandfonline.com/loi/nere20

Accounting for previous performance of students by means of growth curves analyses to estimate the size, stability, and consistency of school effects

Anneke Timmermans & Greetje van der Werf

To cite this article: Anneke Timmermans & Greetje van der Werf (2018): Accounting for previous performance of students by means of growth curves analyses to estimate the size, stability, and consistency of school effects, Educational Research and Evaluation, DOI: 10.1080/13803611.2017.1455300

To link to this article: https://doi.org/10.1080/13803611.2017.1455300

Published online: 04 Apr 2018.

EDUCATIONAL RESEARCH AND EVALUATION, 2018
https://doi.org/10.1080/13803611.2017.1455300

Routledge
Taylor & Francis Group

Accounting for previous performance of students by means of growth curves analyses to estimate the size, stability, and consistency of school effects

Anneke Timmermans and Greetje van der Werf

GION Education/Research, University of Groningen, Groningen, The Netherlands

ABSTRACT

The current study explored the size, stability, and consistency of school effects, using two effectiveness indicators: achievements of students at the end of primary school and growth in achievement across three years of schooling. The sample consisted of the scores of 25,269 students on three subjects, taken in Grades 4 to 6 among 3 cohorts in 319 primary schools. The results showed that (a) for students' growth of achievement the relative proportion of variance at the school level seemed larger compared to achievement at the end of primary school; (b) the total variance in growth was substantially smaller compared to variance in achievement at the end of primary school; (c) school effects for growth are less stable across different cohorts than school effects established at a particular moment; and (d) school effects for growth and less consistent across multiple subject domains than school effects as indicated by students' achievement at a particular moment.

KEYWORDS

Value added; gross school effects; multilevel growth curve modelling; primary education

Introduction

School effectiveness research aims to explain why some schools achieve higher student outcomes than other schools. Initially, these explanations were explored by studying characteristics of schools and school organizations. In the beginning of the 1990s, the focus shifted towards the teacher level and the teaching processes in classrooms. Comprehensive educational effectiveness models, combining factors at several levels among which the school and classroom levels, were developed (e.g., Creemers, 1992; Creemers & Kyriakides, 2008; Scheerens, 1992; Stringfield & Slavin, 1992). These models currently form the conceptual basis of the empirical educational effectiveness studies searching for factors that account for differences in student outcomes between schools and that are malleable, so that they could – in principle – be used for school improvement.

Despite the current model-driven focus of empirical educational effectiveness studies, until now only very small treatment effects on student achievement have been found (Hendriks, 2014; Scheerens, this issue). Moreover, the effects of processes and practices vary considerably across subject matter area and grade level. On the other hand, the

CONTACT Anneke Timmermans A.C.Timmermans@rug.nl

overall or cumulative effects of schooling are by no means small (Scheerens, this issue). However, still no consensus exists about how large the "real" school differences are. Effect sizes of school effects appear to differ between studies, depending on subject matter and grades, but more importantly depending also on whether and which student characteristics were taken into account to compute the "net" school effects (Marks, 2015).

Also, some methodological reservations are in place. First, in many educational effectiveness studies, intelligence or proxies for intelligence like prior achievement are not taken into account; therefore, these studies overestimate the "real" variation in effectiveness between schools in a reliable and valid way (Organisation for Economic Co-operation and Development, 2008; Sammons, Thomas, & Mortimore, 1997). The absence of such controls at the student level may lead to an overestimation of the size of the "real" value added of schools (Meyer, 1997; Timmermans, Doolaard, & De Wolf, 2011; Webster, Mendro, Orsak, & Weerasinghe, 1996). Detterman (2016) states that there is sufficient evidence that in developed countries only 10% of the variance in school achievement can be attributed to schools and teachers, while the remaining 90% is due to student characteristics, of which intelligence is the most important one.

It might also be the case that in most previous educational effectiveness studies the effects of schooling have been underestimated because they only included the achievements of students measured at one specific moment during their school career in the effectiveness indicator. Even those studies that claim to use the "value-added" approach generally are limited, because they have only taken one measurement of prior achievement into account by including it as a co-variate in the analyses. Only a few studies thus far have addressed the value-added approach by analysing differences in effectiveness between schools using the learning gain of students over a longer period of time by means of growth curve modelling. Surprisingly, the studies that did so (e.g., Dumay, Coe, & Anumendem, 2014; Guldemond & Bosker, 2009; Raudenbush, 1989; Rowan, Correnti, & Miller, 2002) were able to demonstrate more sizable school effects. Several types of explanations for higher effect sizes are suggested in these studies, including statistical or conceptual differences in approach (growth versus co-variance analysis) as well as explanations related to school policy characteristics. We will come back to this later in the discussion section.

Second, most previous studies used as an effectiveness indicator only the outcomes of one single cohort of students, mostly pertaining to only one subject domain, which highly threatened the reliability and validity of this indicator. In our view, it is extremely important that when labelling schools as more or less effective, the degree to which they attain higher or lower learning gains among their students across multiple subject domains and across successive student cohorts is also taken into account. On the basis of this approach, it would be possible to rank order schools on one integrated, multidimensional indicator of effectiveness, which offers new perspectives for relating between-school differences in value added to malleable educational process characteristics.

Taking this view as a starting point, it is the aim of the current study to explore the size, stability, and consistency of school effects, using two different effectiveness indicators: the achievements of students at the end of primary school and the learning gains of these students across three years of schooling. The empirical data used in the study are from a Dutch data set including the scores of 25,269 students on reading comprehension, spelling, and mathematics tests, taken in Grades 4, 5, and 6 among three cohorts of students

in 319 primary schools. Because the study builds further on earlier studies reporting on the size, stability, and consistency of school effects and on the statistical models that have been employed in these studies, we will first shortly review the state of the art concerning these issues, after which we will describe more in detail the research questions of the present study.

Size, stability, and consistency of school effects

In a recent publication, Marks (2015) gives a short, but comprehensive overview of relevant research on the size of school effects. Starting with a description of different types of school effects, referring to Scheerens and Bosker (1997), he reviews the findings related to the size of effects in the gross effect models and in the models in which prior differences between students (value-added models) have been adjusted for. Generally, the findings show that the between-school differences in all levels of schooling are much lower in the value-added models (8–13% of the total variance in student academic performance) than in the gross effect models (18–25%). Moreover, the effects are larger for mathematics and composite measures of academic performance compared to language, and larger for schools in secondary education compared to schools in primary education.

However, the sizes of school effects that Marks reported are based on studies that typically used only two measurements of students' achievement, that is, taking the score on the second occasion as the outcome variable, adjusting for the score on the first occasion as an indicator of initial achievement. Furthermore, the reported comparisons of school effect sizes between domains, grades, and school levels are not based on studies conducted in the same schools, and thus it is unknown to which degree the differences that were found could be attributed to fluctuations in the samples of participating schools or within-school fluctuations due to variation between student cohorts (Marks, 2015).

Regarding the stability, Marks (2015) makes a distinction between stability across cohorts and stability across grades within cohorts, although he does not further review the literature on the latter type of stability because this is not an issue in his own study. His conclusion from the review of literature, referring to the study of Scheerens, Bosker, and Creemers (2001) amongst others, is that the stability of value-added school effects is at best moderate, with rather large differences between studies depending on the time lag between the cohorts, with smaller levels of stability the longer the time lag (Gray, Goldstein, & Thomas, 2001; Leckie & Goldstein, 2009; Thomas, Sammons, Mortimore, & Smees, 1997b). For example, Scheerens et al. (2001) presented an average correlation of 0.70 between school effects of subsequent cohorts, but the range of correlations was .34 to .87. The conclusion of moderate stability is in agreement with results of previous studies within the Dutch context (Luyten, 1994; Timmermans, De Wolf, Bosker, & Doolaard, 2015; Van de Grift, 2009).

However, estimates of stability of school effects are difficult to compare over studies as different approaches were adopted in previous research. First, some studies are based on separate value-added analyses for each cohort, retrieval of the value-added estimates from the separate analyses, and calculations of correlations between the value-added estimates (e.g., Marks, 2015). Others are based on separate value-added analyses for each cohort, retrieval of the value-added estimates from the separate analyses, and subsequent

analyses of schools changing categories (e.g., Gray, Jesson, Goldstein, Hedger, & Rasbash, 1995; Van de Grift, 2009). The third category of studies is based on a simultaneous analysis of subsequent cohorts by means of including an additional hierarchical level to the multilevel model to account for dependence of cohorts within schools and then calculating a stability index based on the variance on the school and school-cohort levels (e.g., Leckie & Goldstein, 2009; Luyten, 1994; Van der Werf & Guldemond, 1996). Furthermore, stability or a lack thereof may be interpreted in different ways, that is, it may indicate true changes in effectiveness of schools or departments but it may also be an indication of problems with the reliability of value-added estimates (e.g., Dumay et al., 2014; Inspectie van het Onderwijs, 2003; Wijnstra, Ouwens, & Béguin, 2003; Thomas, Sammons, Mortimore, & Smees, 1997a; Van de Grift, 2009) .

Regarding the consistency of school effects across multiple subject domains, Marks's (2015) conclusion is rather similar: The adjusted correlations between subjects in the studies that he has reviewed (e.g., Hill & Rowe, 1996; Luyten, 1998; Scheerens & Bosker, 1997; Thomas et al., 1997b) are around .50, but again considerable differences in correlations were revealed between studies. The strength of the consistency was dependent on the type of school effect measured, with higher consistency for gross school effects (approximately $r = .70–.80$) than for value-added school effects ($r = .24–.71$). This conclusion is in accordance with the results from previous studies in the Dutch context which are not included in the review of Marks (e.g., Bosker & Luyten, 2000; Timmermans, 2012; Van der Werf & Guldemond, 1996). Altogether, we might conclude that schools which are successful in one year for one subject, are not necessarily successful in the next year or another subject, but it also implies that, generally, good results in one year or one subject to a certain extent go together with good results in another year or another subject (Luyten, 2003).

In his own study, Marks (2015) examined the size, stability, and consistency of school effects for both primary and secondary school students, using population data for Victorian government schools collected among six cohorts of students that were in Year 3 (primary school) or Year 7 (secondary school) between 2008 and 2011. From each cohort, data of two measurement occasions were available: Year 3 and 5 and Year 7 and 9, respectively. Five achievement domains were included in the study. The main conclusions from the study were that, in contrast with the current understanding, value-added school effects show low levels of stability across cohorts and are not consistent across subject domains. With respect to stability, the correlations found by Marks seem to be lower than in previous research, with correlations varying between .21 and .41 for value-added effects of primary schools of 1 year apart. Correlations were very close to zero for value-added effects of primary schools of 2 years apart. For secondary education, the correlations were slightly higher. However, average correlations between subjects varied between .37 and .44 for primary school value added and between .34 and .43 for secondary school value added. These are well within the range of correlations of previous studies. Unfortunately, Marks had only two measurement occasions in the longitudinal data in his study, so he could not report about the size, stability, and consistency of value-added school effects for growth in students achievement; instead, similar to the traditional value-added educational effectiveness studies, he included the achievement scores of the first measurement occasion as a co-variate in the analyses (see the next subsection for an explanation of the difference).

The current study

Our study aims to extend the study of Marks (2015) by adding two important elements. First, because our data set included three measurement occasions (Grades 4, 5, and 6) for each cohort, we estimated the size, stability, and consistency of school effects for students' final scores in Grade 6 (gross school effects) as well as their growth from Grade 4 to 6 (value-added school effects). The three measurement occasions offered the possibility to explicitly model progress over time. It is important to note here that the interpretation of value-added effects based on growth curve models, as applied in the current study, is different from value added derived from the traditional covariance models. The value-added school effect derived from a growth curve model indicates the difference in average growth of the students in school J compared to the average school in the sample. In covariance models, value-added school effects can be interpreted as the difference in the average student performance between school J and the average school for children with comparable prior performance (Timmermans et al., 2011; Willms, 1992). The major difference is that value added derived from traditional covariance models does not explicitly model growth in performance. Second, we integrated the separate models for estimating the size of school effects respectively the stability and consistency into one comprehensive model. The following research questions were addressed:

(1) What is the size of school effects for students' final achievements (gross effects) and their growth (value-added effects) on reading comprehension, spelling, and mathematics?
(2) What is the stability of school effects for the final achievements and growth of students on reading comprehension, spelling, and mathematics across three cohorts of students?
(3) What is the consistency of school effects for the final achievement and growth of students across the three domains of reading comprehension, spelling, and mathematics?

Method

Sample

For this study, data from a sample of Dutch primary schools were derived from the Monitoring and Evaluation System of Cito, the Netherlands Institute for Educational Measurement. This monitoring system offers schools and teachers the possibility to monitor the progress of their students during primary education via several instruments, such as a set of tests, a registration system, remediation guidance methods, and tools for identifying specific learning problems. Data were collected by the schools themselves for their own use.

The data in our study included the results on reading comprehension tests, spelling tests, and mathematics (geometry, time, and money) tests. Three cohorts of students who completed tests in these domains during the period from 2003 to 2008 were included in this study. Students in each cohort were followed for three consecutive years (from Grade 4 until Grade 6, age approximately 9–12 years). Cohort 1 consisted of students in

Grade 4 in the school year 2003–2004 and was followed until Grade 6 in the school year 2005–2006. Similarly, Cohort 2 consisted of students in Grade 4 in 2004–2005 who were followed until Grade 6 in 2006–2007. And finally, the third cohort consisted of students in Grade 4 in 2005–2006 followed until Grade 6 in 2007–2008. Sample sizes for each of the cohorts are presented in Table 1. Within each cohort, there is a gradual decrease of number of students with the largest decrease between Grades 5 and 6. Over cohorts, there seems a gradual increase of student numbers.

Instruments and variables

Student performance

The data set includes student academic performance scores in the domains mathematics, reading comprehension, and spelling. For each of these outcomes, grade-specific tests were developed by Cito, the Netherlands Institute for Educational Measurement. The reliability rates of the grade-specific tests are high. The reliability rates of the tests vary between .83 and .93, indicating high internal consistency (Janssen, Verhelst, Engelen, & Scheltens, 2010; Weekers, Groenen, Kleintjes, & Feenstra, 2011). Within each domain, the students' test results on each particular grade-specific test were calibrated to the other grade-specific tests by means of item response models, more specifically, by means of the one parameter logistic model assuming a one-dimensional underlying latent scale per domain (Verhelst, Glas, & Verstralen, 1993). The result of this calibration is that per domain the scores across all grades are situated on one underlying scale, so that it is possible to determine students' growth over time. Also, the latent scores within a given domain are consistent across the different cohorts. For this particular study, the latent scale scores were standardized over the three cohorts, which allows for comparing the students' final achievement scores and growth across the three cohorts.

Time

A time variable was constructed with the values −2 (Grade 4), −1 (Grade 5), and 0 (Grade 6). Similar operationalizations of the time variables have been used in, for example, the studies of Dumay et al. (2014), Timmermans et al. (2015).

Table 1. Sample sizes for the three cohorts for each test.

		Cohort 1	Cohort 2	Cohort 3
Reading comprehension				
Grade	4	7,468	8,329	8,454
	5	6,927	7,662	7,446
	6	5,065	5,572	5,587
Spelling				
Grade	4	7,434	8,194	8,345
	5	6,881	7,481	7,681
	6	5,063	5,427	5,338
Mathematics				
Grade	4	5,934	6,032	6,173
	5	5,084	5,172	5,312
	6	3,547	3,582	3,638

Analytic strategy

School effects were indexed in a raw form (gross school effect) by means of the students' performance in Grade 6 and in a net form (value-added school effect) measured by the growth of students from Grade 4 until Grade 6. Both indicators of school effects were estimated within the same multivariate multilevel growth curve model using the MLwiN 3.0 software (Rasbash, Steele, Browne, & Goldstein, 2009). Multilevel models are considered the most appropriate because they take the hierarchical structure, and therefore dependency of the data, into account (Snijders & Bosker, 2012). Since multilevel models do not require a strictly balanced design (Quené & Van den Bergh, 2004), all students for whom at least one outcome of one measurement occasion was available were included in the model.

The model employed is a combination of (a) growth curve modelling (Stoel & Galindo Garre, 2011) to account for the previous performance of students and to explicitly estimate growth in student performance; (b) models to test stability of school effects (Leckie, 2013; Luyten, 1994; Van der Werf & Guldemond, 1996); (c) models to test the consistency of school effects (Luyten, 1998; Ma, 2001). In this model, the measurements (Level 1) were nested within students (Level 2), which were nested in school cohorts (Level 3), which were nested in schools (Level 4). A description of the characteristics of this model will be provided below (the full mathematical model is described in Appendix 1).

First, in multilevel growth curve models the development of the students was modelled as a function of time (Stoel & Galindo Garre, 2011) by including a hierarchical level of measurements (Level 1) within students (Level 2). Time was added as a predictor variable at the measurement level. Given that in the current study only three measurement occasions were available, only linear trends of time were estimated, to prevent an overestimation of between-student, school-cohort, and school differences due to measurement error at particular time points. Gross and value-added school effects can be derived from this model because the time variable in the current study was constructed with the values −2 (Grade 4), −1 (Grade 5), and 0 (Grade 6). This makes Grade 6 the reference time point. Random intercepts on the measurement, student, school-cohort, and school levels were added to the model to allow for the separation of the variance in Grade 6 performance over the hierarchical levels. They refer to the gross effects. Also, random slopes of time were added to the model on the student, school-cohort, and school levels to allow for differences between students, school cohorts, and schools in growth rates. These random slopes refer in our study to value-added effects.

Second, the stability of school effects can be investigated by including multiple cohorts in one analysis and having the cohort level as a separate level within schools (Leckie, 2013; Luyten, 1994; Van der Werf & Guldemond, 1996). In the current model, the school cohorts (Level 3) are nested within schools (Level 4). The variance in the model associated with the school level, both with respect to the intercepts (gross effects) and the slopes (value-added effects), gives an indication of the enduring effects, while the variance on the school-cohort level provides an indication of the instability.

Third, the consistency of the school effects (Research question 3) can be derived from a multilevel model by including multiple dependent variables (Luyten, 1998; Ma, 2001). This makes the multilevel model a multivariate model. In the current study, three dependent variables are considered, that is reading comprehension, spelling, and mathematics. The gross

and value-added effects are estimated for each of the three dependent variables in the model. In the random part of the model, the random intercepts and random slopes of the different dependent variables are allowed to be associated. Therefore, the gross and value-added school effects are allowed to be related within domains as well as across domains. Consistency can therefore be determined for both types of school effects as well as the stable (school-level) and unstable (school-cohort-level) part of the school effects.

The aim of the current study is to estimate the size of school effects for the three outcomes, including (growth of and final performance on) reading comprehension, spelling, and mathematics (Research question 1), the stability of school effects for the three outcomes (Research question 2), and consistency of school effects (Research question 3). The answers on the first research question can for the gross effects be derived from this multilevel model by calculating intra-class correlations (ICC; Snijders & Bosker, 2012). ICC is the common effect size indicator in educational effectiveness research to establish the contribution of schooling in correlational studies. The size of the stable gross school effects (over the three cohorts) can be estimated as the proportion of school-level intercept variance (Level 4) relative to the total intercept variance. The size of the unstable school effects can be estimated as the proportion of the combined school and school-cohort level intercept variance (Levels 3 and 4) relative to the total intercept variance. The latter should give similar results as when estimating the school-level variance in a single-cohort study. No consensus has yet been achieved on how to calculate an estimate of the size of (stable and unstable) school effects for growth (value-added school effects) from growth models or similar statistical techniques. Moreover, the term ICC is not used as an indicator of between-school differences in achievement growth. In this paper, we follow the guidelines as presented by Dumay and colleagues (2014) by calculating and presenting "the school-level slope variance as a proportion of the total slope variance" (p. 70)[1]; a similar approach was adopted in Rowan et al. (2002) and Anumendem, De Fraine, Onghena, and Van Damme (2017). Several other studies present values for between-school differences for growth rates in multilevel growth models without presenting exact descriptions of how they were calculated or providing insufficient information to reconstruct the formulas used (see Guldemond & Bosker, 2009; Raudenbush, 1989). In addition to the intra-class correlations (for gross effects) and proportion of school-level variance (for value-added effects), we also present 50% and 95% coverage intervals to facilitate understanding of between-school differences for both the gross and value-added school effects.

Stability indexes (Research question 2) were also derived from the variance components of the multilevel model. The stability of school effects is estimated as the proportion of stable school effects (Level 4) relative to the sum of stable and unstable school effects (Levels 3 and 4) (Leckie, 2013; Luyten, 1994, 2003). These estimates of stability of school effects can be calculated for both gross school effects and value-added school effects in the exact same way. Consistency (Research question 3) can simply be derived from the estimated covariance in the random part of the model.

The model presented in the results is based on 66,815 measurements of 25,269 students in 868 school cohorts in 319 primary schools. This implies that, on average, data of 2.64 out of the maximum 3 measurements are available per student, data from 29.11 students are available per school cohort, and data from on average 2.72 cohorts out of

a maximum of 3 cohorts are available per school. This indicates that the data are rather complete.

Results

Descriptive statistics of the three cohort samples

The correlation tables for each of the cohorts are presented in Table 2. The strongest correlations are found between the subsequent measures of the same subject domains, of which the highest were found for reading comprehension ($r = .680–.736$) followed by spelling ($r = .655–.707$) and mathematics ($r = .591–.705$). The correlations between the measures for the three different subject domains are also considerable ($r = .305–.578$), with the smallest correlation between spelling Grade 4 and mathematics Grade 6 in Cohort 2 and the largest correlation between reading comprehension Grade 6 and mathematics Grade 6 in Cohort 3. These between-outcome associations confirm the necessity for a multivariate model in which the three outcomes are modelled simultaneously.

Table 3 shows the mean and standard deviations of the scores per grade and per cohort. It seems that for all three subject domains the results per grade are almost equal across cohorts and the growth in scores across grades indicates a linear trend, although only three measurement occasions per cohort are actually too few to reliably estimate this.

Table 2. Correlations between test scores.

		Reading comprehension			Spelling			Mathematics		
	Grade	4	5	6	4	5	6	4	5	6
Cohort 1										
Reading comprehension	4	1								
	5	.729**	1							
	6	.697**	.734**	1						
Spelling	4	.412**	.427**	.419**	1					
	5	.383**	.419**	.422**	.707**	1				
	6	.418**	.447**	.460**	.655**	.667**	1			
Mathematics	4	.475**	.461**	.470**	.368**	.342**	.348**	1		
	5	.513**	.547**	.559**	.371**	.365**	.393**	.653**	1	
	6	.480**	.490**	.569**	.344**	.328**	.382**	.599**	.705**	1
Cohort 2										
Reading comprehension	4	1								
	5	.712**	1							
	6	.694**	.728**	1						
Spelling	4	.384**	.388**	.409**	1					
	5	.366**	.388**	.405**	.698**	1				
	6	.399**	.408**	.469**	.664**	.656**	1			
Mathematics	4	.485**	.466**	.492**	.366**	.337**	.347**	1		
	5	.500**	.532**	.566**	.338**	.338**	.373**	.660**	1	
	6	.463**	.480**	.555**	.305**	.312**	.357**	.592**	.704**	1
Cohort 3										
Reading comprehension	4	1								
	5	.699**	1							
	6	.680**	.736**	1						
Spelling	4	.382**	.397**	.422**	1					
	5	.346**	.409**	.415**	.693**	1				
	6	.401**	.433**	.474**	.655**	.661**	1			
Mathematics	4	.486**	.449**	.459**	.353**	.311**	.342**	1		
	5	.496**	.530**	.542**	.337**	.334**	.359**	.672**	1	
	6	.442**	.493**	.578**	.306**	.331**	.391**	.591**	.678**	1

*$p < .01$. **$p < .001$.

Table 3. Means and standard deviations testscores, per grade and per cohort.

		Reading comprehension			Mathematics			Spelling		
Cohort		Grade 4	Grade 5	Grade 6	Grade 4	Grade 5	Grade 6	Grade 4	Grade 5	Grade 6
1	*M*	39.28	49.30	58.84	62.78	76.18	88.34	135.64	141.70	148.28
	N	7,468	6,927	5,065	5,934	5,084	3,547	7,434	6,881	5,063
	SD	14.91	15.48	16.16	11.54	12.74	12.82	7.39	6.51	6.35
2	*M*	39.72	49.02	58.47	62.91	75.98	88.61	135.47	141.76	148.15
	N	8,329	7,662	5,572	6,032	5,172	3,582	8,194	7,481	5,427
	SD	15.05	15.44	15.92	11.49	12.78	12.75	7.36	6.44	5.89
3	*M*	39.53	49.83	59.53	62.50	75.55	89.49	135.71	141.89	148.49
	N	8,454	7,446	5,587	6,173	5,312	3,638	8,345	7,681	5,257
	SD	14.91	15.17	15.95	11.75	12.76	12.70	7.30	6.43	6.20

Results from the multilevel analysis

The results from the multivariate multilevel growth curve model are presented in Table 4 for the fixed part and in Table 5 for the random part. The fixed part of the model only consists of the intercepts for the three outcomes, the Grade 6 sample average over the three cohorts, and the coefficients for the linear effect of the time variable. On average, students in Grade 6 scored 0.59 points on reading comprehension, 0.82 on spelling, and 0.88 on mathematics. These are scores on the underlying latent scale, that was standardized per outcome variable over all three cohorts. Furthermore, the slopes indicate that from Grade 4 to Grade 6 the students gained on average 0.55 points per year on reading comprehension, 0.75 points per year on spelling, and 0.79 points per year on mathematics. The growth for spelling and mathematics is therefore on average approximately three quarters of a standard deviation per year, against a bit more than half a standard deviation for reading comprehension.

The random effects in Table 5 are presented in variance-covariance matrices for each level included in the multilevel model. The variances of the intercepts (gross effects) and slopes (value-added effects) are presented on the diagonal and the covariances between them off-diagonal. The results from Table 5 indicate that there is a significant amount of variance for each of the outcome variables on each of the hierarchical levels in Grade 6 performance (intercept variance), as well as significant differences in growth between units at all levels (slope variance). The school-level intercept variance (gross school effect) for spelling, for example, is .043, which is significantly different from zero; $z = 8.6$, $p < .001$. With respect to the variance components, it is important to note that the variance in slopes (value added) is smaller than the variance in intercepts (gross effects). Below, we will discuss the results on the school level and the school-cohort level more in detail.

Table 4. Results from the fixed-effects part of the multivariate multilevel growth curve model.

	Estimate	*SE*	CI 2.5%	CI 97.5%	*p* value
Intercept reading comprehension	0.590	0.017	0.556	0.624	<.001
Intercept spelling	0.816	0.015	0.788	0.845	<.001
Intercept mathermatics	0.878	0.020	0.839	0.916	<.001
Time reading comprehension	0.553	0.006	0.540	0.565	<.001
Time spelling	0.745	0.008	0.730	0.760	<.001
Time mathematics	0.790	0.008	0.774	0.806	<.001

Table 5. Results from the random-effects part of the multivariate multilevel growth curve model (variance covariance matrix per hierarchical level).

	RC (int.)		SP (int.)		Math (int.)		RC (slp.)		SP (slp.)		Math (slp.)	
Level: School												
RC (int.)	0.067	0.008										
SP (int.)	0.034	0.005	0.043	0.005								
Math (int.)	0.057	0.007	0.030	0.005	0.075	0.009						
RC (slp.)	0.006	0.002	0.003	0.002	0.004	0.002	0.006	0.001				
SP (slp.)	0.001	0.002	0.010	0.002	−0.001	0.003	0.001	0.001	0.011	0.002		
Math (slp.)	0.009	0.003	0.005	0.002	0.015	0.003	0.002	0.001	0.001	0.001	0.005	0.002
Level: School cohort												
RC (int.)	0.026	0.004										
SP (int.)	0.004	0.002	0.028	0.003								
Math (int.)	0.013	0.003	0.000	0.003	0.031	0.004						
RC (slp.)	0.013	0.002	0.002	0.001	0.006	0.002	0.009	0.001				
SP (slp.)	0.001	0.001	0.015	0.002	0.001	0.002	0.001	0.001	0.010	0.001		
Math (slp.)	0.006	0.002	0.001	0.002	0.021	0.003	0.004	0.001	0.002	0.001	0.021	0.002
Level: Student												
RC (int.)	0.595	0.008										
SP (int.)	0.287	0.005	0.303	0.005								
Math (int.)	0.398	0.007	0.222	0.005	0.453	0.008						
RC (slp.)	0.036	0.003	0.027	0.002	0.041	0.002	0.007	0.002				
SP (slp.)	−0.015	0.002	−0.049	0.002	−0.003	0.002	−0.002	0.001	0.002	0.001		
Math (slp.)	0.044	0.003	0.021	0.002	0.044	0.003	0.01	0.001	0.006	0.001	0.008	0.002
Level: Measurement												
RC (int.)	0.212	0.002										
SP (int.)	0.006	0.001	0.186	0.002								
Math (int.)	0.012	0.002	0.000	0.002	0.176	0.002						

Note: Standard errors are given in parenthesis. RC = reading comprehension; SP = spelling; Math = mathematics; int. = intercept variance; slp. = slope variance.

Size of the stable school effects (Research question 1)

As was explained earlier, the variance in the model associated with the school level, both with respect to the intercepts (gross effects) and the slopes (value-added effects), gives an indication of the stable school effects, while the variance on the school-cohort level provides an indication of the unstable school effects.

The variances on the school level indicate that schools differ in both their stable gross effects and in their stable value-added effects. In order to illustrate the size of school effects intra-class correlations, 50% and 95% coverage intervals are presented in Table 6. Relative to the total variance in Grade 6 performance, the proportion of variance on the school level (gross school effect) is .074 for reading comprehension[2], .077 for spelling, and .102 for Mathematics, respectively. The 95% coverage intervals show that for Grade 6 performance 95% of the schools vary approximately between 0.083 and 1.097 points on reading comprehension, between 0.410 and 1.222 for spelling, and between 0.341 and 1.415 for mathematics.[3] These differences between schools are considerable as the differences between stronger and weaker schools are almost one standard deviation in reading comprehension and spelling and a bit over one standard deviation for mathematics. The 50% coverage intervals are much narrower, indicating that when the 50% schools in the middle range are considered, differences are much less pronounced.

Compared to the gross school effects, there is a relatively larger proportion associated with the school level for the growth between Grade 4 and Grade 6 (value-added school effects); intra-class correlations are .147 for mathematics, .273 for reading comprehension[4], and .478 for spelling, respectively. For the value-added effects, 95% of schools vary in growth rates between 0.401 and 0.705 standard deviations for reading comprehension, between 0.539 and 0.951 standard deviations for spelling, and between 0.651 and 0.929 standard deviations per year for mathematics. Again, when the middle 50% of the distribution of schools is considered, differences in value added are much smaller. The difference between a school on the 25th percentile and the 75th percentile for the three subjects is approximately one tenth of a standard deviation.

Table 6. Estimates of the size and stability of school effects for both Type 0 and Type A school effects.

		Reading comprehension	Spelling	Mathematics
Size stable school effects				
Gross effect (intercept)	ICC	.074	.077	.102
	50% CI	[0.416, 0.764][1]	[0.676, 0.956]	[0.693, 1.063]
	95% CI	[0.083, 1.097]	[0.410, 1.222]	[0.341, 1.415]
Value added (slope)	Between-school differences[2]	.273	.478	.147
	50% CI	[0.501, 0.605][3]	[0.674, 0.816]	[0.742, 0.838]
	95% CI	[0.401, 0.705]	[0.539, 0.951]	[0.651, 0.929]
Size unstable school effects				
Gross effect (intercept)	ICC	.103	.127	.144
	50% CI	[0.384, 0.796]	[0.636, 0.996]	[0.659, 1.097]
	95% CI	[−0.008, 1.188]	[0.294, 1.338]	[0.240, 1.516]
Value added (slope)	Between-school differences	.682	.913	.765
	50% CI	[0.470, 0.636]	[0.647, 0.843]	[0.681, 0.899]
	95% CI	[0.313, 0.739]	[0.461, 1.029]	[0.474, 1.106]
Stability index				
Gross effect (intercept)	ICC	.720	.606	.708
Value added (slope)	ICC	.400	.524	.192

[1]50 and 95% coverage intervals are centred around the estimated intercepts (see Table 3). [2]Between-school differences as proposed by Dumay et al. (2014), based on the random slopes of the time variable on the student, school-cohort, and school levels. [3]50 and 95% coverage intervals are centred around the estimated slopes (see Table 3).

Size of the unstable part of school effects (Research question 1)

Also within schools, there is substantial variation between the three subsequent school cohorts for the gross and the value-added school effects, which indicates that the effects of school are partially unstable. Relative to the total variance in Grade 6 performance (gross effects), the proportion of variance on the combined school and school-cohort level is .103 for reading comprehension, .127 for spelling, and .144 for mathematics, respectively. With respect to the growth between Grade 4 and Grade 6 (value added), there is a relatively larger proportion of variance associated with the combined school and school-cohort level; .682, .913, and .765, respectively. The proportion of the combined variance on the school and school-cohort levels is relevant as it has a similar meaning as the proportion of the school-level variance in a single-cohort study.

The gross effects of 95% of the school cohorts vary approximately between −0.008 and 1.188 points on the standardized reading comprehension scale, between 0.294 and 1.338 for spelling, and between 0.240 and 1.516 for mathematics. This implies that for all three outcomes the differences between stronger and weaker school cohorts in gross effects are larger than one standard deviation. The differences in the middle range are again less pronounced. The gross effects of 50% of the school cohorts vary approximately between 0.384 and 0.796 points on the standardized reading comprehension scale, between 0.636 and 0.996 for spelling, and between 0.659 and 1.097 for mathematics. For the value-added effects, 95% of the school cohorts vary in growth rates between 0.313 and 0.739 standard deviations for reading comprehension, between 0.461 and 1.029 standard deviations for spelling, and between 0.474 and 1.106 standard deviations per year for mathematics. These findings imply substantial differences in growth rates between cohorts within schools. The difference in value added between a school on the 25th percentile and the 75th percentile for the three subjects is approximately two tenths of a standard deviation when the unstable school effects are considered.

Stability of gross and value-added school effects (Research question 2)

Table 6 also includes stability indexes, that is, the proportion of the school-level variance relative to the combined school and school-cohort level variance (Leckie, 2013). The gross school effects, referring to the average performance of students in Grade 6, are relatively stable as their indexed stability values show: .720 for reading comprehension, .606 for spelling, and .708 for mathematics. This means that the stable part of the gross school effects is larger than the unstable part of the gross school effects. For the value-added school effects, it is the other way around: The unstable part of the school effect tends to be larger than the stable part, with stability indexes of .400 for reading comprehension, .524 for spelling, and .192 for mathematics, respectively. It therefore seems that the gross school effects are relatively more stable than the value-added school effects.

Consistency of school effects (Research question 3)

The consistency of the stable and unstable school effects is derived from the random part of the model by the correlations of the school effects for different outcomes (see Table 7). Correlations ranging from .52 (mathematics and spelling) to .80 (mathematics

Table 7. Estimates of consistency of school effect through correlations.

Gross effect(intercept)	Reading comprehension	Spelling	Mathematics
School level			
Reading comprehension	1		
Spelling	0.64	1	
Mathematics	0.80	0.52	1
Value added (Slope)			
Reading comprehension	1		
Spelling	0.15	1	
Mathematics	0.44	0.12	1
School-cohort level			
Gross effect (intercept)			
Reading comprehension	1		
Spelling	0.14	1	
Mathematics	0.45	−0.01	1
Value added (Slope)			
Reading comprehension	1		
Spelling	0.10	1	
Mathematics	0.30	0.13	1

and reading comprehension) were found for the gross effects on the school level. This indicates that the stable gross school effects are relatively consistent over outcomes. Schools with relatively high means on one outcome tend to have high means on the other outcomes as well. For the stable value-added school effects, the correlations ranged between .15 and .44. These positive, but much weaker, correlations indicate that schools in which students make on average more progress in one outcome tend to realize more progress in the other outcomes as well. However, these correlations are far from perfect. At the school-cohort level, the correlations are generally positive as well, but relatively low, both for the gross school effects ($r = -.01–.45$) as for the value-added effects ($r = .10–.30$). This indicates that the unstable part of the school effects is less consistent than the stable part; this holds for both gross school effects and value-added effects.

Discussion

Taking into account the stability and consistency of school effects, is, in our view, extremely important in order to be able to distinguish in a reliable and valid way between relatively effective and less effective schools. It would give a more detailed picture of the "real" value added of schools and would do more justice to the complex nature of schooling (Timmermans, 2012). The aim of the current study was to test whether these assumptions could be supported by empirical data of a Dutch data set, which includes the scores on reading comprehension, spelling, and mathematics tests, taken in Grades 4, 5, and 6 among three cohorts of students in primary schools. Our study extended the study of Marks (2015), with two important elements. First, our data set included three measurement occasions (Grades 4, 5, and 6) for three cohorts, which made it possible to estimate growth of students' achievement over time. Second, we integrated the separate models for estimating the size of school effects respectively the stability and consistency into one comprehensive model.

Size of school effects

The findings of the current study indicate smaller between-school differences with respect to the gross school effects (stable gross school effects 7–10% between-school variance, ICC varying between .074 to .102; see Table 6, upper part, first row) compared to previous research (18–25%). In accordance to previous research, the school effects in this study were slightly bigger for mathematics compared to language. The above-mentioned difference in size of school effects between previous studies and ours can partly be attributed to differences in methodology. By including several cohorts, the variation due to stable school effects and year-to-year differences could be disentangled. The part of the variances that can be accounted for by stable school effects is smaller than the year-to-year school effects (10–14% between-school-cohort variance, intra-class correlations varying between .103 and .144; see Table 6, second part, first row), which indicated that estimates from single-cohort studies tend to overestimate school effects.

The findings of previous research indicate that the between-school differences in all levels of schooling are much lower in the value-added models (8–13%) compared to gross school effects. However, in the current study the differences between schools in the progress of students (value-added school effects) were relatively large (15–48%, proportions of slope variance between .147 and .478; see Table 5, first part, second row), again with higher levels of year-to-year differences in progress (Table 5, second part, second row) than stable value-added school effects. With respect to value-added school effects, remarkably, we found smaller between-school differences for mathematics (15% between-school variance in the stable part) compared to languages (27 and 48% for reading comprehension and spelling, respectively). The results from the current study stem from multilevel growth models, which allow for an explicit modelling of the progress made by students as a function of time instead of taking the final performance and correct that for initial performance of students. Previous research has indicated that growth models are able to demonstrate more sizable school effects (e.g., Dumay et al., 2014; Guldemond & Bosker, 2009). How this might be explained is not yet totally clear.

Conceptually, the growth model (and also the learning gain model) differs from the commonly used co-variance model. In the latter model, student performance differences between schools at a particular moment in time are predicted, taking into account the differences in prior performance of the students of these schools. Schools which score higher than was predicted from their students prior performance are considered as more effective than schools which score lower than the predicted performance. In the growth (and learning gain) approach, however, the differences between schools in how much students have learned over time are predicted. Schools in which students have learned more or quicker over time are considered as more effective than schools in which students have learnt less over the same period of time (see also Veenstra, 1999). However, also models analysing learning gain instead of performance status at a particular moment in general yielded much smaller school effects than our and other researchers' growth models. According to Rowan et al. (2002), this is due to the fact that annual learning gain scores (the score difference between two measurement moments) are very problematic, because they could be very unreliable when true differences among students in academic growth are small. This probably explains why the effect sizes yielded by learning gain models generally are very small. As an alternative for covariate adjustment models

and learning gain models, Rowan et al. suggest to analyse the effects of teachers (or schools) on students' achievement growth by using a cross-classified random effects model, as proposed by Raudenbusch and Bryk (2002). They expect that this model, just as other growth models like ours, allows to separate true score variance in growth rates from error variance, and thus to ignore error variance in determining the effects at the different levels of analysis. Following this line of reasoning, this improves the estimates of teacher effects over those derived from simple gain models, and this in turn might lead to higher sizes of school effects. And indeed, in their own study, Rowan et al. found that the size of the teacher effect was 3 times larger than the size of the effects in other studies. So, for the time being, we might conclude that the different and presumably more reliable way of multilevel growth modelling explains the differences in effect sizes that we found in our study in comparison with other – simple learning gain – studies. The explanation for the difference with other studies in which just co-variate adjustment was applied actually is not relevant, because conceptually this type of studies in fact do not address the issue of value added of schools in terms of change of student performance over time.

However, also an additional elaboration on the effect sizes used in this and previous studies is in place. Most used effect-size indicators in educational effectiveness research, among which ICC, only provide us with an indication of the proportion of the unexplained variance at the different hierarchical levels, for example, the proportion of unexplained variance at the school level. In the current study, however, the total variation in slopes (value added) is quite limited compared to the variation in intercepts (gross effects). In such a case, comparing between the gross and value-added school effects may be suboptimal. In the current study, this finding may indicate that most variation in students' academic achievement is already present at the first measurement occasions (between schools and within schools). Therefore one may conclude that, although the distribution of variance is significantly different when it comes to intercepts and slopes, the variation in value added is relatively small and unlikely to alter the initial situation very much. This is in line with numerous studies on educational effectiveness which show that variation in achievement (between schools and between students) is largely determined by student background (e.g., initial achievement, IQ, socioecomic status [SES]) and only to a limited extent by educational factors (like quality of instruction, time on task, etc.).

Stability and consistency

With respect to the stability of school effects, we see higher levels of stability indexes for the gross school effects compared to the value-added school effects. Stability indexes well over .50 indicate that for the gross school effects the stable school effects were larger than the year-to-year differences. For the value-added estimates, the stability indexes of reading comprehension and mathematics which were lower than .50 suggest that the year-to-year differences are larger than the stable school value-added effects. Only for spelling, the stable school effects were slightly bigger than the year-to-year differences. These findings confirm that increased sophistication in modelling learning gains seems to be associated with decreased stability of the estimated school effects over time (e.g., Dumay et al., 2014; Gray et al., 2001; Thomas, Peng, & Gray, 2007). There may be several

reasons why value-added school effects show lower stability than gross school effects, such as attenuation of correlations due to restriction of range and decreased reliability due to statistical adjustments (Dumay et al., 2014); or school effects may actually not be as stable as many researchers believed (Marks, 2015), or schools may strive for a more or less stable target in final performance (Luyten & De Wolf, 2011). There is an important task for future research to try and explain the relatively low stability of value-added estimates of school effects derived from more sophisticated statistical models.

On the basis of previous research, Marks (2015) concluded that the strength of the consistency was dependent on the type of school effect measured, with higher consistency for gross school effects (approximately $r = .70–.80$) than for value-added school effects ($r = .24–.71$). The current study confirms these findings. The current study adds to the existing literature that consistency and stability are interrelated, as the consistency of school effects is stronger for the stable school effects compared to the year-to-year differences. The latter result suggests that schools with a highly stable gross effect in a particular subject tend to have a highly stable gross effect in other subjects. It also indicates that within schools there are differences in effects for subsequent cohorts of students, but cohorts that have high performance or progress in one subject tend to do well in the other subjects as well. Combining the statistical approaches for estimating consistency and stability revealed that correlations indicating consistency from single-cohort studies may be affected by both the stable effects of schools as well as the specific student cohort.

Limitations and suggestions for future studies

Limitations of the current study should be kept in mind when interpreting the findings. The most important limitation might be the absence of potential important covariates (e.g., ethnicity, SES, language status, etc.) that could have been incorporated into the model in order to test their respective sensitivity to them. However, in the study of Marks (2015) including SES had a neglectable effect on top of prior achievement. Also, because our models addressed growth of student achievement over time, it might well be that SES was implicitly taken into account as well, because it is highly correlated with prior achievement of students.

Furthermore, by estimating the multilevel growth models we assume that differences in growth between schools are valid indicators of the contribution that each school makes to the cognitive growth of students. The validity of such indicators is hard to establish in research. However, research on summer learning (Downey, von Hippel, & Broh, 2004; Heyns, 1978) indicates that most of the variation in growth occurs during the summer vacation. In the current study, there was no information on student performance in the beginning of the school year, right after the summer holidays. Therefore, we could not test the variation in summer vacation growth differences. Also, because our data set contained too many students with a deviating school career, because they started when they were younger or older than 4 years old at the (former) official deadline of 1 October, we could not distinguish the actual schooling effects from age effects. Future research on the same data set is recommended to unravel the effects of individual student characteristics on growth of performance from the effects of particular school populations or cohorts.

Finally, in the present case, growth has been modelled as a linear function of time. In other circumstances, for example, when more measurements are available or when a longer time period is covered, a quadratic function or even more complex forms of modelling the development of achievement over time would be more appropriate. In those cases, it would also be possible to model declining growth. Modelling more growth using more complex, non-linear models would make it much more complicated to apply the approach presented in this paper. In the present situation, linear growth can be expressed by a single statistic (growth per year). In the case of non-linear growth, this becomes more complicated. For example, if a quadratic function is used to model declining growth, the variances of both the linear and quadratic effects need to be taken into account. Perhaps this complication can be avoided by using log transformations of test scores, but in that case the interpretation of the results will be less straightforward.

Conclusions

Taken together, the results of our study demonstrated some important findings which might have implications for future school effectiveness research. First, we showed, just as some previous studies did, that at first sight school effects seem larger for students' growth of achievement than for achievement at a particular point in time, taking into account prior achievement at an earlier moment. However, the total variation in slopes (value added) is quite limited compared to the variation in intercepts (gross effects), which may lead to the conclusion that, although the distribution of variance is significantly different when it comes to intercepts and slopes, the variation in value added is relatively small and unlikely to alter the initial situation very much. We also demonstrated that this type of school effect is less stable across different cohorts of students than school effects established at a particular point in time. Also, school effects as indicated by students' achievement growth over time are less consistent across multiple subject domains than school effects as indicated by students' achievement at a particular moment. Therefore, it is important for future research that only the stable component of these school effects as well as the consistency across subject domains are included in the effectiveness indicators. In combination with including students' growth of achievement over time, and with unravelling age effects from schooling effects, this might lead to a more reliable and valid value-added effectiveness indicator, with also a higher proportion of variance that could possibly be explained by malleable school and teacher characteristics. However, also some other issues have to be resolved beforehand. In the discussion of our findings, we did an attempt to explain how it would be possible that our results deviated considerably from the results of other educational effectiveness studies in which more traditional models were applied. Nevertheless, it still remains unclear how the differences in results could be explained, because not only different statistical models were applied in the studies, but the studies were also different regarding design, measurement instruments, number of measurement moments, age groups, and contexts (countries). Therefore, following Dumay et al. (2014), we would like to recommend for future research to compare different methods of estimating the size, stability, and consistency of school effects; preferably a comparison of methods on one data set. As an example, we did a re-analysis of our Grade 4 and Grade 6 data, applying co-variance adjustment modelling as well as modelling the gain scores between Grades 4 and 6. The results can be found

in Appendix 2. Generally, they show that the school effects in the gross and value-added co-variance models are quite equal to the results of similar models in other studies. The results of the gain scores models, however, show that the stable and unstable school effects are larger compared to the value-added co-variance models, but smaller than in the growth models presented in the main part of this article. With respect to stability, the results show moderate to high stability indices, with the lowest value for the value-added effect models. However, for both models the indices are higher than in the earlier presented growth models. Regarding consistency, the findings are more or less similar for both models presented in Appendix 2, and the consistency appeared to be higher than was found in the growth models.

In conclusion, these additional analyses show that school effects indeed differ according to the use of different statistical modelling approaches, and that these differences are, at least partly, related to conceptual differences among approaches. However, we consider the results of this article, including the additional analyses, only as a first step, which hopefully challenges educational effectiveness researchers and multilevel experts to bring more conceptual and methodological clearness regarding the issues that were raised by our findings.

Notes

1. The application of the formulas as proposed by Dumay and colleagues (2014) implies that the intra-class correlations of the gross effects and values for the school effects for growth cannot be compared directly, as the variance component of the measurement occasion level is included in the estimation of the ICC for the gross school effect (included variance components all relate to intercept variance), but not in the computation of the value-added school effect (all variance components relate to slope variance).
2. Calculated, for example, for reading comprehension, as the proportion of intercept variance in reading comprehension at the school level (0.067) in relation to the total intercept variance (0.067 + 0.026 + 0.595 + 0.212 = 0.90; see Table 5). The resulting intraclass correlation is 0.074 (= 0.067/0.90).
3. The derivation of 95% coverage intervals is based on the model assumption that the random effects are normally distributed. Given normality, we expect 95% of the random effects for each level to lie in the range of ± 1.96 times the square root of the associated variance component. When reporting coverage intervals, it is often helpful to centre them around the intercept or some other interpretable value (Leckie, 2013, p. 18).
4. For the value-added effect of reading comprehension, for example, the value for the between-school difference in growth is calculated as the proportion of school-level slope variance (0.006) in relation to the total slope variance (0.006 + 0.009 + 0.007 = 0.022; see Table 5). This proportion is 0.273 (0.006/0.022).

Acknowledgement

This research did not receive any specific grants from funding agencies in the public, commercial, or not-for-profit sectors.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Anumendem, N. D., De Fraine B., Onghena, P., & Van Damme, J. (2017). Growth in reading comprehension and mathematics achievement in primary school: A bivariate transition multilevel growth curve model approach. *Biometrics & Biostatistics International Journal*, *5*(4): 00137. doi:10.15406/bbij.2017.05.00137

Bosker, R. J., & Luyten, H. (2000). De stabiliteit en consistentie van differentiële schooleffecten [The stability and consistency of differential school effects]. *Tijdschrift voor Onderwijsresearch*, *24*(3/4), 308–321.

Creemers, B. P. M. (1992). School effectiveness, effective instruction and school improvement in the Netherlands. In D. Reynolds & P. Cuttance (Eds.), *School effectiveness, research, policy and practice* (pp. 48–70). London: Cassell.

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

Detterman, D. K. (2016). Education and intelligence: Pity the poor teacher because student characteristics are more significant than teachers or schools. *The Spanish Journal of Psychology*, *19*(E93), 1–11. doi:10.1017/sjp.2016.88

Downey, D. B., von Hippel, P. T., & Broh, B. A. (2004). Are schools the great equalizer? Cognitive inequality during the summer months and the school year. *American Sociological Review*, *69*(5), 613–635. doi:10.1177/000312240406900501

Dumay, X., Coe, R., & Anumendem, N. D. (2014). Stability over time of different methods of estimating school performance. *School Effectiveness and School Improvement*, *25*(1), 64–82. doi:10.1080/09243453.2012.759599

Gray, J., Jesson, D., Goldstein, H., Hedger, K., & Rasbash, J. (1995). A multilevel analysis of school improvement: Changes in schools' performance over time. *School Effectiveness and School Improvement*, *6*(2), 97–114. doi10.1080/0924345950060201

Gray, J., Goldstein, H., & Thomas, S. (2001). Predicting the future: The role of past performance in determining trends in institutional effectiveness at A level. *British Educational Research Journal*, *27*(4), 391–405. doi:10.1080/01411920125622

Guldemond, H., & Bosker, R. J. (2009). School effects on students' progress – A dynamic perspective. *School Effectiveness and School Improvement*, *20*(2), 255–268. doi:10.1080/09243450902883938

Hendriks, M. A. (2014). *The influence of school size, leadership, evaluation, and time on student outcomes* (Doctoral dissertation). Enschede: University of Twente.

Heyns, B. (1978). *Summer learning and the effects of schooling*. New York, NY: Academic Press.

Hill, P. W, & Rowe, K. J. (1996). Multilevel modelling in school effectiveness research. *School Effectiveness and School Improvement*, *7*(1), 1–34. doi:10.1080/0924345960070101

Inspectie van het Onderwijs. (2003). *Het bepalen van de toegevoegde waarde door basisscholen* [To assess the value added of primary schools]. Utrecht: Author.

Janssen, J., Verhelst, N., Engelen, R., & Scheltens, F. (2010). *Wetenschappelijke verantwoording van de toetsen LOVS Rekenen-Wiskunde voor groep 3 tot en met 8* [Scientific validation of the mathematics tests from the monitoring system Grade 1 until Grade 6]. Arnhem: Cito.

Leckie, G. (2013). *Three-level multilevel models: Concepts. LEMMA VLE Module 11*. Bristol: Centre for Multilevel Modelling, University of Bristol.

Leckie, G., & Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *172*(4), 835–851. doi:10.1111/j.1467-985X.2009.00597

Luyten, H. (1994). Stability of school effects in Dutch secondary education: The impact of variance across subjects and years. *International Journal of Educational Research*, *21*(2), 197–216. doi:10.1016/0883-0355(94)90032-9

Luyten, H. (1998). School effectiveness and student achievement, consistent across subjects? Evidence from Dutch elementary and secondary education. *Educational Research and Evaluation*, *4*(4), 281–306. doi:10.1076/edre.4.4.281.6950

Luyten, H. (2003). The size of schools effects compared to teacher effects: An overview of the research literature. *School Effectiveness and School Improvement*, *14*(1), 31–51. doi:10.1076/sesi.14.1.31.13865

Luyten, H., & De Wolf, I. (2011). Changes in student populations and average test scores of Dutch primary schools. *School Effectiveness and School Improvement*, *22*(4), 439–460. doi:10.1080/09243453.2011.591614

Ma, X. (2001). Stability of school academic performance across subject areas. *Journal of Educational Measurement*, *38*(1), 1–18. doi:10.1111/j.1745-3984.2001.tb01114.x

Marks, G. N. (2015). The size, stability, and consistency of school effects: Evidence from Victoria. *School Effectiveness and School Improvement*, *16*(3), 397–414. doi:10.1080/09243453.2014.964264

Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review*, *16*(3), 283–301. doi:10.1016/S0272-7757(96)00081-7

Organisation for Economic Co-operation and Development. (2008). *Measuring improvements in learning outcomes: Best practices to assess the value-added of schools*. Paris: Author.

Quené, H., & Van den Bergh, H. (2004). On multi-level modelling of data from repeated measures designs: A tutorial. *Speech Communication*, *43*(1–2), 103–121. doi:10.1016/j.specom.2004.02.004

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2009). *A user's guide to MLwiN, Version 2.10*. Bristol: Centre for Multilevel Modelling, University of Bristol.

Raudenbush, S. W. (1989). The analysis of longitudinal, multilevel data. *International Journal of Educational Research*, *13*(7), 721–740. doi:10.1016/0883-0355(89)90024-4

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis* (2nd ed.). Thousand Oaks, CA: Sage.

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. *Teacher College Record*, *104*(8), 1525–1567. doi:10.1111/1467-9620.00212

Sammons, P., Thomas, S., & Mortimore, P. (1997). *Forging links: Effective schools and effective departments*. London: Paul Chapman.

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

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

Scheerens, J., Bosker, R. J., & Creemers, B. P. M. (2001). Time for self-criticism: On the viability of school effectiveness research. *School Effectiveness and School Improvement*, *12*(1), 131–157. doi:10.1076/sesi.12.1.131.3464

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modelling* (2nd ed.). Los Angeles, CA: Sage.

Stoel, R. D., & Galindo Garre, F. (2011). Growth curve analysis using multilevel regression and structural equation modeling. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 97–111). New York, NY: Routledge.

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. 35–69). Groningen: ICO.

Thomas, S., Sammons, P., Mortimore, P., & Smees, R. (1997a). Differential secondary school effectiveness: Comparing the performance of different pupil groups. *British Educational Research Journal*, *23*(4), 451–469. doi:10.1080/0141192970230405

Thomas, S., Sammons, P., Mortimore, P., & Smees, R. (1997b). Stability and consistency in secondary schools' effects on students' GCSE outcomes over three years. *School Effectiveness and School Improvement*, *8*(2) , 169–197. doi:10.1080/0924345970080201

Thomas, S., Peng, W. J., & Gray, J. (2007). Modelling patterns of improvement over time: Value added trends in English secondary school performance across ten cohorts. *Oxford Review of Education*, *33*(3), 261–295. doi:10.1080/03054980701366116

Timmermans, A. C. (2012). *Value added in educational accountability: Possible, fair and useful?* (Doctoral dissertation). Groningen: GION Onderwijs/Onderzoek.

Timmermans, A. C., De Wolf, I. F., Bosker, R. J., & Doolaard, S.(2015). Risk-based educational accountability in Dutch primary education. *Educational Assessment Evaluation and Accountability*, *27*(4), 323–346. doi:10.1007/s11092-015-9212-y

Timmermans, A. C., Doolaard, S., & De Wolf, I. (2011). Conceptual and empirical differences among various value-added models for accountability. *School Effectiveness and School Improvement*, *22*(4), 393–413. doi:10.1080/09243453.2011.590704

Van de Grift, W. (2009). Reliability and validity in measuring the value added of schools. *School Effectiveness and School Improvement*, *20*(2), 269–285. doi:10.1080/09243450902883946

Van der Werf, M. P. C., & Guldemond, H. (1996). *Omvang, stabiliteit en consistentie van schooleffecten in het basisonderwijs* [Size, stability, and consistency of school effects in primary education]. Groningen: GION, Gronings Instituut voor Onderzoek van Onderwijs, Opvoeding en Ontwikkeling, Rijksuniversiteit Groningen.

Veenstra, R. (1999). *Leerlingen-klassen-scholen: Prestaties en vorderingen van leerlingen in het voortgezet onderwijs* [Students-classes-schools: Achievement and progress of students in secondary education]. Groningen: ICS (Interuniversity Centre for Social Science Theory and Methodology).

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1993). *OPLM: One parameter logistic model. Computer program and manual*. Arnhem: Cito.

Webster, W. J., Mendro, R. L., Orsak, T. H., & Weerasinghe, D. (1996, April). *The applicability of selected regression and hierarchical linear models to the estimation of school and teacher effects*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Weekers, A., Groenen, I., Kleintjes, F., & Feenstra, H. (2011). *Wetenschappelijke verantwoording papieren toetsen Begrijpend lezen voor groep 7 en 8* [Scientific validation of the paper tests for reading comprehension Grade 5 and Grade 6]. Arnhem: Cito.

Wijnstra, J., Ouwens, M., & Béguin, A. (2003). *De toegevoegde waarde van de basisschool* [The value added of primary schools]. Arnhem: Cito.

Willms, J. D. (1992). *Monitoring school performance: A guide for educators*. Washington, DC: The Falmer Press.

Appendix 1. Mathematical properties of the estimated model

The following formulas represent the mathematics properties of the multivariate four-level hierarchical model.

Reading comprehension

$$Y_{1ijkl} = \beta_{01jkl} + \beta_{11jkl} Time_{ijkl} + e_{1ijkl}$$

$$\beta_{01jkl} = \gamma_{001} + w_{01l} + v_{01kl} + u_{01jkl}$$

$$\beta_{11jkl} = \beta_{011} + w_{11l} + v_{11kl} + u_{11jkl}$$

Spelling

$$Y_{2ijkl} = \beta_{02jkl} + \beta_{12jkl} Time_{ijkl} + e_{2ijkl}$$

$$\beta_{02jkl} = \gamma_{002} + w_{02l} + v_{02kl} + u_{02jkl}$$

$$\beta_{12jkl} = \beta_{012} + w_{12l} + v_{12kl} + u_{12jkl}$$

Mathematics

$$Y_{3ijkl} = \beta_{03jkl} + \beta_{13jkl} Time_{ijkl} + e_{3ijkl}$$

$$\beta_{03jkl} = \gamma_{003} + w_{03l} + v_{03kl} + u_{03jkl}$$

$$\beta_{11jkl} = \beta_{013} + w_{13l} + v_{13kl} + u_{13jkl}$$

Specification of the random effects

$$\begin{pmatrix} w_{01l} \\ w_{02l} \\ w_{03l} \\ w_{11l} \\ w_{12l} \\ w_{13l} \end{pmatrix} \sim \left\{ N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{w01} & & & & & \\ \sigma^2_{w01,02} & \sigma^2_{w02} & & & & \\ \sigma^2_{w01,03} & \sigma^2_{w02,03} & \sigma^2_{w03} & & & \\ \sigma^2_{w01,11} & \sigma^2_{w02,11} & \sigma^2_{w03,11} & \sigma^2_{w11} & & \\ \sigma^2_{w01,12} & \sigma^2_{w02,12} & \sigma^2_{w03,12} & \sigma^2_{w11,12} & \sigma^2_{w12} & \\ \sigma^2_{w01,13} & \sigma^2_{w02,13} & \sigma^2_{w03,13} & \sigma^2_{w11,13} & \sigma^2_{w12,13} & \sigma^2_{w13} \end{pmatrix} \right) \right\}$$

$$\begin{pmatrix} v_{01kl} \\ v_{02kl} \\ v_{03kl} \\ v_{11kl} \\ v_{12kl} \\ v_{13kl} \end{pmatrix} \sim \left\{ N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{v01} & & & & & \\ \sigma^2_{v01,02} & \sigma^2_{v02} & & & & \\ \sigma^2_{v01,03} & \sigma^2_{v02,03} & \sigma^2_{v03} & & & \\ \sigma^2_{v01,11} & \sigma^2_{v02,11} & \sigma^2_{v03,11} & \sigma^2_{v11} & & \\ \sigma^2_{v01,12} & \sigma^2_{v02,12} & \sigma^2_{v03,12} & \sigma^2_{v11,v12} & \sigma^2_{v12} & \\ \sigma^2_{v01,13} & \sigma^2_{v02,13} & \sigma^2_{v03,13} & \sigma^2_{v11,v13} & \sigma^2_{v12,v13} & \sigma^2_{v13} \end{pmatrix} \right) \right\}$$

$$\begin{pmatrix} u_{01jkl} \\ u_{02jkl} \\ u_{03jkl} \\ u_{11jkl} \\ u_{12jkl} \\ u_{13jkl} \end{pmatrix} \sim \left\{ N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u01} & & & & & \\ \sigma^2_{u01,02} & \sigma^2_{u02} & & & & \\ \sigma^2_{u01,03} & \sigma^2_{u02,03} & \sigma^2_{u03} & & & \\ \sigma^2_{u01,11} & \sigma^2_{u02,11} & \sigma^2_{u03,11} & \sigma^2_{u11} & & \\ \sigma^2_{u01,12} & \sigma^2_{u02,12} & \sigma^2_{u03,12} & \sigma^2_{u11,12} & \sigma^2_{u12} & \\ \sigma^2_{u01,13} & \sigma^2_{u02,13} & \sigma^2_{u03,13} & \sigma^2_{u11,13} & \sigma^2_{u12,13} & \sigma^2_{u13} \end{pmatrix} \right) \right\}$$

$$\begin{pmatrix} e_{1ijkl} \\ e_{2ijkl} \\ e_{3ijkl} \end{pmatrix} \sim \left\{ N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{e1} & & \\ \sigma^2_{e1,2} & \sigma^2_{e2} & \\ \sigma^2_{e1,3} & \sigma^2_{e2,3} & \sigma^2_{e3} \end{pmatrix} \right) \right\}$$

In this model, measurements (indexed by subscript *i*) at Level 1 are nested within students (*j*) at Level 2, who are nested within school cohorts (*k*) at Level 3, which are nested within schools (*l*) at Level 4. For each of the three dependent variables (Reading comprehension Y_{1ijkl}, Spelling Y_{2ijkl}, Mathematics Y_{3ijkl}), a separate regression formula is fitted.

For each dependent variable, the regression formula fitted consists of an intercept and the slope of time. Time was added as a predictor variable at the measurement level. Given that in the current study only three measurement occasions were available, only linear trends of time were estimated, assuming a single pace in development over the two school years. The time variable in the current study was constructed with the values −2 (Grade 4), −1 (Grade 5), and 0 (Grade 6), by which the Grade 6 performance of students became the reference value. The intercepts and slopes of time were fitted for each dependent variable separately, by which we allow the three variables to have their own mean intercept and slope for time (Table 4). Or in other words, the Grade 6 performance and growth between Grade 4 and Grade 6 in reading comprehension was allowed to deviate from spelling and mathematics.

Both the intercepts and slopes are allowed to be random at the higher levels. This implies that students, school cohorts, and schools are allowed to have their own intercepts and slopes of time. With respect to gross school effects, deviations from the grand mean in Grade 6 performance (Reading comprehension γ_{001}, Spelling γ_{002}, Mathematics γ_{003}) were allowed for schools (Reading comprehension w_{01l}, Spelling w_{02l}, Mathematics w_{03l}), school cohorts within schools (Reading comprehension v_{01kl}, Spelling v_{02kl}, Mathematics v_{03kl}), and students within the cohorts (Reading comprehension u_{01jkl}, Spelling u_{02jkl}, Mathematics u_{03jkl}). For the value-added effects, deviations from the average slope of time (Reading comprehension β_{011}, Spelling β_{012}, Mathematics β_{013}) were allowed for schools (Reading comprehension w_{11l}, Spelling w_{12l}, Mathematics w_{13l}), cohorts within schools (Reading comprehension v_{11kl}, Spelling v_{12kl}, Mathematics v_{13kl}), and students within cohorts (Reading comprehension u_{11jkl}, Spelling u_{12jkl}, Mathematics u_{13jkl}).

The random effects are assumed to have each a normal distribution with a mean of 0 and one standard deviation to be estimated. They are also assumed to be statistically independent over

Table A1. Results of the three additional models.

Gross effects model				Value added effects model				Gain score model			
Fixed effects				**Fixed effects**				**Fixed effects**			
	B	*SE*(B)			B	*SE*(B)			B	*SE*(B)	
Intercept RC	58.35	0.33		Intercept RC	58.58	0.23		Intercept RC	18.47	0.32	
Intercept SP	148.17	0.13		Intercept SP	14.82	0.12		Intercept SP	12.20	0.19	
Intercept Math	88.13	0.32		Intercept Math	88.03	0.26		Intercept Math	24.27	0.41	
				RC grade 4	0.69	0.01					
				SP grade 4	0.54	0.01					
				Math grade 4	0.62	0.01					
Random effects				**Random effects**				**Random effects**			
	RC	SP	Math		RC	SP	Math		RC	SP	Math
School level											
RC	20.54 (2.54)			RC	7.82 (1.24)			RC	19.38 (2.35)		
SP	4.89 (0.79)	2.74 (0.38)		SP	1.14 (0.45)	2.03 (0.30)		SP	8.40 (1.16)	7.70 (0.85)	
Math	13.54 (2.02)	3.82 (0.72)	14.62 (2.19)	Math	3.21 (1.06)	1.41 (0.51)	5.91 (1.45)	Math	17.31 (2.47)	11.91 (1.52)	26.88 (3.72)
School-cohort level											
RC	6.21 (1.12)			RC	7.08 (0.89)			RC	8.86 (1.05)		
SP	0.10 (0.37)	1.56 (0.22)		SP	0.71 (0.31)	1.80 (0.19)		SP	1.45 (0.39)	2.48 (0.27)	
Math	2.84 (0.93)	−0.24 (0.39)	7.133 (1.16)	Math	3.44 (0.89)	0.50 (0.40)	11.50 (1.39)	Math	5.25 (1.11)	1.89 (0.55)	16.79 (1.86)
Student level											
RC	232.96 (2.62)			RC	122.22 (1.38)			RC	128.98 (1.45)		
SP	42.31 (0.80)	34.0 9 (0.39)		SP	8.77 (0.40)	17.96 (0.21)		SP	1.15 (0.49)	25.49 (0.29)	
Math	103.79 (1.90)	27.69 (0.70)	144.84 (1.97)	Math	31.02 (1.07)	7.11 (0.41)	89.63 (1.24)	Math	16.49 (1.10)	2.74 (0.49)	92.67 (12.64)

the levels. However, within each hierarchical level the random effects are allowed to correlate. This implies that at each level a full variance-covariance matrix is estimated (Table 5).

Appendix 2. Traditional school effect models applied to the same dataset

In addition to the multilevel multivariate growth curve model, we estimated a series of alternative models based on the same original data. However, due to model specifications the number of schools, school cohorts, and students included in the three models differ from the growth curve model. In these models, the number of units on each of these levels is reduced considerably. Comparisons between these models and the model from the main article should therefore be handled with caution.

All additional models include the school-cohort level in order to investigate stability and are multivariate to deal with the three domains (see Method section). The first model is a gross school effect model in which the Grade 6 performance on the three domains served as dependent variables, and no explanatory variables are included in the model. The second model is a value-added effects model in which the Grade 6 performance on the three domains again served as dependent variables. In this model, we added Grade 4 performance as explanatory variables. Specifically, Grade 4 performance for reading comprehension was added as explanatory for Grade 6 reading comprehension, but not for the dependent variables in the other domains. Similar strategies were applied for the other Grade 4 performance scores. The third model is a gain score model in which for each domain the Grade 4 scores were subtracted from the Grade 6 scores. The gain scores then served as dependent variables in the model. No explanatory variables were included in the third model. The outcomes of these models are presented in Table A1.

For each of the three additional models, an estimation of between-school differences was provided, calculated as the proportion school-level variance in relation to the total variance in the dependent variable (sum of school, school-cohort, and student-level variance). These are presented separately for each domain in the first part of Table A2.

In the gross school effect, the proportion of variance on the school and school-cohort levels is relatively small, indicating that most of the variance in Grade 6 performance is located at the student level. The proportions found in these models match with results from previous research in which similar models were employed. The proportions of variance at the school and school-cohort level derived from the gain score model seem larger compared to the two other models. For all three models, we see a significant proportion of variance at the school-cohort level, which implies that estimations of between-school differences based on single cohorts are likely an overestimation of the stable differences in school effects.

Table A2. Estimated school effects, stability and consistency.

	Gross effects model			Value-added effects model			Gain score model		
Between-school differences	RC	SP	Math	RC	SP	Math	RC	SP	Math
School level	0.08	0.07	0.09	0.06	0.09	0.06	0.15	0.22	0.20
School-cohort level	0.10	0.11	0.13	0.11	0.18	0.16	0.18	0.29	0.32
Stability index	0.77	0.64	0.67	0.52	0.53	0.34	0.69	0.69	0.73
Consistency									
School level									
RC	1			1			1		
SP	0.64	1		0.29	1		0.69	1	
Math	0.78	0.60	1	0.47	0.41	1	0.76	0.83	1
School-cohort level									
RC	1			1			1		
SP	0.03	1		0.19	1		0.31	1	
Math	0.43	-0.07	1	0.38	0.11	1	0.43	0.29	1

Stability indexes were estimated in a similar fashion as in the main article as a proportion of the school-level variance in relation to the sum of the school and school-cohort variance (see Method section); see second part Table A2. With respect to the stability, the results show moderate to high stability indices, with the lowest for the value-added effects model.

Concerning consistency over domains, correlations between variance components at the school and school-cohort level are presented in the third part of Table A2. With respect to consistency, the results of the gross effects and gain score model are similar. Both models produce school-effect estimates that are relatively consistent over the three domains. The correlations between the school-level variance components from the value-added effects model are considerably smaller. At the school-cohort level, the consistency is, for each of the three domains, considerably smaller.