School composition on student proficiencies

The effect of student background and school composition on the static scores and gain scores of cognitive and non-cognitive proficiency tests in Dutch primary schools

Rusconi, R.; Volman, M.L.L.; van der Ark, L.A.

DOI
10.2139/ssrn.4316516

Publication date
2023

Document Version
Final published version

SCHOOL COMPOSITION ON STUDENT PROFICIENCIES

The effect of student background and school composition on the static scores and gain scores of cognitive and non-cognitive proficiency tests in Dutch primary schools

Ralph Rusconi[1,2], Monique L. L. Volman[2], and L. Andries van der Ark[2]

[1]Department of Education, Municipality of Amsterdam, The Netherlands

[2]Research Institute of Child Development and Education, University of Amsterdam, The Netherlands

Author note

Ralph Rusconi, corresponding author https://orcid.org/0000-0002-8324-3184

Monique L. L. Volman https://orcid.org/0000-0001-9217-1402

L. Andries van der Ark https://orcid.org/0000-0003-3131-7943

Correspondence concerning this article should be sent to Ralph Rusconi, Dept of Education Municipality of Amsterdam, , PO Box 1840, 1000 BV Amsterdam, Email: r.rusconi@amsterdam.nl, tel nr: +31 624212827.

We have no conflict of interest to disclose.

Acknowledgements and supplemental materials

We would like to thank Sjoerd Karsten for his help conceptualizing the study. Furthermore we would like to thank Lissette Hornstra for her help with using MLwinN. The underlying supplemental material (S1), the ML-analyses of the models used for this article, can be accessed at osf.io/r53ts .

SCHOOL COMPOSITION ON STUDENT PROFICIENCIES

The effect of student background and school composition on the static scores and gain scores of cognitive and non-cognitive proficiency tests in Dutch primary schools

Abstract

Schools are often evaluated based on student proficiency. Evaluating schools, as little as possible distorted by external factors is preferable. We investigate whether student background characteristics and school composition have a different effect on cognitive and non-cognitive student proficiencies in value-added models using gain scores compared to static scores. To investigate which models are preferable Data of 4,918 students at 216 primary schools in the Netherlands were used. Their static scores and gain scores in mathematic proficiency, reading comprehension, and the non-cognitive proficiencies of task motivation and self-efficacy were measured. The effects of student background and school composition were estimated. Student background and school composition had a larger effect on static scores than gain scores, and these effects are stronger for cognitive than for non-cognitive proficiencies, with student background having a larger effect on both proficiencies than school composition. The results offer new arguments for researchers to use gain scores.

Keywords: *evaluation method; mathematic proficiency; non-cognitive proficiency; reading comprehension; school composition; school evaluation; value-added model.*

1. Introduction

There has been a growing emphasis on performance in education, both at the school and system level. Student-proficiency data are used to monitor the progress of students, and the quality of teachers and schools. The use of and decisions based on these data are becoming more and more imperative (Braun, 2005; Dronkers & Robert, 2008; Kelly & Majerus, 2011).

Student proficiency encompasses both cognitive and non-cognitive proficiencies. Important cognitive proficiencies include proficiencies in the teaching language and in mathematics as they form the basis of understanding all other subjects in school (Houtveen, van de Grift, & Creemers, 2004). Non-cognitive proficiencies deemed important in a successful school career include self-efficacy, motivation and well-being. All three have a significant effect on proficiency in the cognitive development and, in their own right, are considered skills required to succeed in life and a future career (Durlak et al., 2011; OECD, 2015). Cognitive proficiencies are correlated with student background characteristics (e.g., SES and ethnicity) and school characteristics, such as school composition, quality of teachers, school organization, and school leadership (Driessen, 2002; Van Der Slik, Driessen, & De Bot, 2006).

Cognitive proficiencies are typically measured using tests, and non-cognitive proficiencies using questionnaires. For the evaluation of teachers and schools, three types of methods correcting the test or questionnaire scores are currently in use: no correction, correcting for student background characteristics, and correcting for student and school characteristics. Today, the evaluation without taking into account student backgrounds or school composition is used only in non-scientific settings; for example, evaluating schools based purely on school rankings (e.g., League Tables, Goldstein & Cuttance, 1988).

As of the 1960s, correcting scores for student background became more common in teacher and school evaluations (e.g., Coleman, 1966). This is still a common method when evaluating schools by student proficiency (Faubert, 2009). As of the 1990s, teacher and school evaluations started correcting for both student background and school characteristics (e.g., Van Der Slik et al., 2006). Scheerens (2012), among others, argued that it is important to hold teachers and schools accountable, but only for the part that they actually can control. Teachers and schools typically have no control over student background characteristics and certain school characteristics, such as school composition (distribution of the student population in terms of, e.g., SES, ethnicity, and regular v. special-need students), but they may have control over the quality of the school, including the quality of teachers, school organization, and leadership. To evaluate the effect of schools and teachers on student proficiency, only the school-quality indicators should be taken into account. To this end value-added models (VAMs) have been developed (e.g., Bosker, 2012; Harris, 2011b; Saunders, 1999; Strand, 1997; Timmermans, 2012). By correcting for as many relevant background variables as possible, VAMs provide an estimate of the student-proficiency variance that is due to school quality, and thus provide the value added by the school (Braun, 2005; OECD, 2008).

In general, three types of proficiency scores are being used in school evaluation. First, a single score obtained at the end of an evaluation cycle, for example, the score on an end-of-school test. In the remainder, we call these scores *static scores*. The second score is the difference between the scores of comparable tests at the beginning and the end of an evaluation cycle, which are known as *gain scores* (e.g., Timmermans, 2012). The third score is a single score that has been corrected for by an earlier test score. Hence the first test score is used as a covariate and the second test score as an independent variable. In the remainder, we call this a *corrected static score*. Corrected static scores rather than gain scores are used if the test scores at the beginning and end of an evaluation cycle cannot be calibrated on a single scale (Timmermans, 2012). For VAMs, the use of gain scores and corrected static scores are advocated (e.g., Bosker, 2012; Harris, 2011a, 2011b) but static scores are still commonly used in school evaluation (Faubert, 2009; Zvoch & Stevens, 2008).

For static scores, the effect of both student background characteristics and school composition on student proficiency is significant (e.g., McConney & Perry, 2010; Zvoch & Stevens, 2008). For school and teacher evaluations scores should therefore be corrected for both student background characteristics and school composition. For gain scores, the effect of student background on student proficiency is significant (Verhaeghe et al., 2011), but the effect of school composition on gains scores of student proficiency has not been fully investigated, especially in primary education. Hence, it is not fully known whether correction for school composition is necessary.

1.1 Objective of This Study

In this study, we investigated the effects of two correction methods (only student background characteristics, and school composition and student background characteristics) on cognitive and non-cognitive student proficiency using both static and gain scores. In particular, we investigated the effects of SES and ethnicity at the school- and student-level. Findings about the model with the least influenced by the independent variables might give schools and other investigators the tools to strengthen their inference on the school effects on student proficiencies.

2. Review of Related Literature

2.1 Cognitive and non-cognitive student proficiency

Educational effectiveness research is often narrow in its focus on cognitive proficiency (Coe & Fitz-Gibbon, 1998; Timmermans, 2012). In the public debate, in accountability, and in research, school and teacher accountability also focuses mostly on cognitive proficiencies (Bill & Melinda Gates Foundation, 2014; Braun, 2005; Gabriel & Allington, 2012). The focus on cognitive proficiencies can be justified as it is one of the main tasks of a school to develop the students' knowledge and skills in cognitive subjects such as mathematical and language proficiency. Moreover, in the Netherlands mathematic proficiency and reading comprehension are tested throughout a student's career, making these two subjects ideal for school effectiveness studies.

A meta-analysis containing more than 200 studies showed a positive relation between social development and school achievement (Durlak et al., 2011), which justifies an additional focus on non-cognitive development. Several researchers and policy makers (e.g., Coe & Fitz-gibbon, 1998; Konu, Lintonen, & Autio, 2002; OECD, 2015) argued that schools should not teach only cognitive skills, but teach a complete set of abilities including non-cognitive skills. Peetsma, Hascher, Van der Veen, and Roede (2005) showed that non-cognitive student skills, such as well-being and motivation, differ among schools. In the light of school accountability and the broader scope including non-cognitive skills, it is important to know which schools are able to increase gains in the development of non-cognitive outcomes while not sacrificing cognitive gains. Three non-cognitive skills−self-efficacy, task motivation, and well-being−are important factors in cognitive student outcomes and deemed important 'life skills' in their own right (Konu et al., 2002; Opdenakker & Van Damme, 2000; Peetsma et al., 2005). Research shows that schools exercise little control over the static scores of non-cognitive proficiency (M. C. C. Opdenakker & Van Damme, 2000). Timmermans (2012) showed that for secondary schools a positive value-added in the non-cognitive characteristics of *classroom climate* and *achievement motivation* does not always correlate with cognitive proficiencies. As the emphasis on non-cognitive development grows and schools differ in their contribution to the development of these non-cognitive skills, it is relevant to find accurate measures of non-cognitive outcomes in the context of school accountability (Teddlie & Reynolds, 2000).

2.2 Correction method

School composition is defined as the average composition of the student population in terms of ethnicity, socio-economic status (SES), gender, and possibly other relevant covariates, and varies across schools (Driessen, 2002). School composition measured by average SES of the students on the one side and the SES of the student's family on the other side predict student proficiency equally well (Caldas & Bankston, 1998; Rumberger & Palardy, 2005).

School composition by ethnicity also correlates with student proficiency (Bankston & Caldas, 1998; Caldas & Bankston, 1998; Driessen, 2002). The percentage of ethnic minority students in a school correlates negatively with proficiency, although the correlation becomes very small when corrected for SES (Driessen, 2002). Verhaeghe et al. (2011) showed that SES and ethnicity do affect the initial proficiency with which students start their school career, but they found no effect of SES and ethnicity on the value-added of the schools on the students' subsequent proficiency.

Schools can differ in average non-cognitive proficiency of their students and this is related to the school's composition (Bradley & Corwyn, 2002; Palardy, 2008). The effects of school composition on non-cognitive static proficiencies are however small (Konu et al., 2002; M. C. C. Opdenakker & Van Damme, 2000) and significantly smaller than the effects of school composition on cognitive static proficiencies (Opdenakker & Van Damme, 2000). Teddlie and Reynolds (2000) found a correlation between school composition effects on cognitive proficiency and on non-cognitive static proficiencies. This might indicate a double effect; a student is influenced by school composition on both cognitive and non-cognitive proficiencies. This held especially true for primary schools.

Rumberger and Palardy (2005) showed that the effect of school composition on gains in student proficiencies in mathematics, science, reading, and history can be as large as the effect of the student's own background characteristics. This effect was explained by teacher expectations, the amount of homework students do, number of courses that students take, and the students' feelings about safety. School composition by SES can also influence the school environment composition by SES can have a peer group effect on top of differences in climate, pedagogical approach and financing (Driessen, 2007). Schools with a high average SES can expect higher parental contributions, both financial and in terms of investment of time and effort, for example, parents helping with reading in class, than low SES schools (Opdenakker, Damme, deFraine, Van Landeghem, & Onghena, 2002).

2.3 Further notes on the use of gain scores and static scores

A prerequisite of using gains scores is that the two test scores (pre-test, post-test) must be calibrated on a single scale, which is done by a method called equating (Embretson, 1996). If gain scores are unavailable, and two possibly unrelated tests are used, an alternative procedure may be used (Timmermans, 2012; Verhaeghe et al., 2011). The score on the first test can be included as a covariate in the regression, so that the score on the second test is corrected for both student characteristics and the score on the first test. However, if the two tests are not selected carefully and measure different attributes, the correlation between the tests scores may be low, which prohibits making meaningful inferences.

There is a current trend to use VAMs for school or teacher evaluation, preferably using gain scores or corrected static scores, making them high stakes decision making tools (Braun, 2005; Whitehurst, 2013). But modelling gain scores in VAMs poses challenges not encountered in modelling static scores (Hanushek & Rivkin, 2010). Following students over longer time periods may suffer from control variables that change over time, like the SES of parents, and from attrition and the resulting missing-data problem. Hence, the lag between the two test administrations should be considered carefully. If it is too long, the challenges may hamper the investigation, if it is too short, the inferences may be invalid. In addition, the results of VAMs using gain scores may be harder to explain.

3. Research Question

To investigate which model is most suited to evaluate schools based on student proficiencies we focused primarily on the effect of student background and the combined effects of student background and school composition on cognitive and non-cognitive proficiency tests. We expected that student background would have a larger effect on cognitive than on non-cognitive proficiency test scores (e.g., Driessen, 2009; Hofman et al., 1999; Opdenakker & Van Damme, 2000; Van Landeghem et al., 2002, Verhaeghe et al., 2011). We expected that the effect of school composition on cognitive proficiency test scores is smaller than the effect of student background, so the value added of correcting for school composition as well is small (e.g., Van Landeghem et al., 2002; Verhaeghe et al., 2011). Furthermore, we expected that these effects were greater for static scores than for gain scores (Timmermans, 2012). We did not expect interaction effects. Table 1 shows an overview of our expectations based on the literature discussed.

Table 1

Expected Relative Effects of Correction Method, Type of Proficiency, and Scoring Method on Cognitive and Non-Cognitive Proficiency Test Scores.

Scoring method	Correction method			
	Student background		School composition	
	Proficiency		Proficiency	
	Cognitive	Non-cognitive	Cognitive	Non-cognitive
Static score	++++	+++	++	++
Gain score	+++	++	+	+

Note: The number +s indicate the relative strength of the positive effects. For example, +++ indicates a stronger positive effect than ++.

4. Method

4.1 Subjects

Data were obtained from the longitudinal *Cool 5-18* cohort study in the Netherlands (Driessen et al., 2009), in which children in grades K2, 3, and 6 in primary school were tested in Dutch reading comprehension and mathematic proficiency every three years. Students also completed a questionnaire measuring task motivation, well-being, and self-efficacy in grades 3 and 6. (Driessen, 2009; Driessen et al., 2009, 2012). The student background characteristics ethnicity and SES were collected using a parent questionnaire. This large-scale cohort study followed the school attainment of students in Dutch schools between the ages of 5 and 18, in 2008, 2011, and 2014. The data used in this study were collected in 2008 and 2011. The tests in Dutch reading comprehension and mathematic proficiency used in this study are currently still in use in almost all primary schools in the Netherlands (Cito, 2015; Tomesen et al., 2019).

Our sample consisted of 5,960 students who completed the tests in 2008 (in grade 3) and in 2011 (in grade 6). Approximately half of the sample completed a different test version for mathematic proficiency or reading comprehension than the modal test version. These students were dropped from the sample, and the final sample sizes equaled 2,375; 3,052; and 4,118 for reading comprehension, mathematic proficiency, and non-cognitive proficiencies, respectively (Table 2). The response rates exceeded 92% and the non-respondents were randomly distributed over the different SES and ethnicity groups. Approximately 200 schools dropped out of the study between measurements, leaving 342 schools that participated in both measurements. Driessen et al. (2012) found that school attrition was neither correlated with background characteristics of the students nor with their achievement levels. The major reason for student attrition was due to a change of school or class. Approximately 88% of the students were still in the same school at the time of the second measurement in 2011. Attrition of students had no significant correlation with SES, ethnicity, or prior achievement in the first measurement. For details on the data collection, we refer to Driessen et al. (2009, 2012). For a multilevel regression analysis a sample size of 100 is advised at level 2 (schools) and a group size (student level) of 10 students (Maas & Hox, 2005, 2004). Both criteria are met by the data collected in this study (Table 2).

Table 2:
Descriptive Statistics of the Student-Level and School-Level Variables

Dependent variables	Acronym	Students	Schools	Mean	*SD*
Reading comprehension gain	RG	2383	133	30.51	17.69
Reading comprehension static	RS	2383	133	56.39	15.62
Mathematic proficiency gain	MG	3055	168	39.57	17.46
Mathematic proficiency static	MS	3055	168	109.70	12.72
Task motivation gain	TG	4918	261	-0.28	0.869
Task motivation static	TS	4918	261	3.94	0.627
Self-efficacy gain	EG	4918	261	-0.13	0.912
Self-efficacy static	ES	4918	261	3.68	0.624
Well-being gain	WG	4918	261	0.05	0.89
Well-being static	WS	4918	261	3.82	0.58
Independent variables	**Acronym**	**Students**	**Schools**	**Mean**	***SD***
SES level 1 high[a]	SH1	4797	261	0.30	0.46
SES level 1 medium[a]	SM1	4797	261	0.42	0.49
SES level 1 low[ab]	SL1	4797	261	0.27	0.45
SES level 2	SES2	[c]	261	0.90	0.31
Ethnicity level 1[a]	E1	4027	261	0.23	0.42
Ethnicity level 2	E2	[c]	261	0.24	0.32

[a] Dichotomous variables. [b] In the analyses, SES at level 1 was represented by two dummy variables with SES level 1 low as the reference category. [c] Not applicable

4.2 Dependent Variables

Reading comprehension was measured using the test *Begrijpend Lezen M5 and M8 version 1997* (Cito, 2015; Feenstra et al., 2010; Hollenberg et al., 2011); a reading comprehension test series for grades 2 to 6. Mathematic proficiency was measured using the test *Rekenen en Wiskunde M5 and M8 version 2007* (Janssen et al., 2010), a test series measuring proficiency in a range of mathematical subjects across grades 1 through 6. The test scores obtained in the two test administrations were equated, enabling a meaningful comparison between the 2008 and 2011 administrations. For both reading comprehension and mathematical proficiency, the scores obtained in 2011 (Grade 6) were used as static scores. We used the acronyms RS (reading comprehension static) and MS (mathematic proficiency static) to indicate the static scores for reading comprehension and mathematical proficiency, respectively. For both cognitive proficiencies the gain scores used in the VAMs were obtained by the difference of the scores on the grade-3 test and the grade-6 test. We used the acronyms RG (reading comprehension gain) and MG (mathematic proficiency gain) to indicate the gain scores for reading comprehension and mathematical proficiency, respectively.

Task motivation (TS and TG for static scores and gain scores, respectively) was measured using five items of the Goal Orientation Questionnaire (Seegers, van Putten, & de Brabander, 2002), self-efficacy (ES and EG for static scores and gain scores, respectively) was measured using six items from a Dutch version of the Patterns of Adaptive Learning Survey (PALS; Midgley et al., 2000), and well-being (WS and WG for static scores and gain scores, respectively) was measured using a 13-item scale developed by Peetsma, Wagenaar, and De Kat (2002).Table 3 shows the characteristics of the measurement instruments.

Table 3
Descriptives of Measurement Instruments in COOL-cohort Study

Instrument	Attribute	Items	Rel.
Begrijpend lezen M5	Reading comprehension	50[a]	.85
Begrijpend lezen M8	Reading comprehension	50[a]	.84
Rekenen en wiskunde M5	Mathematic proficiency	28[a]	.93
Rekenen en wiskunde M8	Mathematic proficiency	32[a]	.97
Goal Orientation Questionnaire	Task motivation	5[b]	.74
Patterns of Adaptive Learning Survey	Self-efficacy	6[b]	.78
Questionnaire by Peetsma et al. (2002)	Well-being	13[b]	.78

Note. Rel = Reliability estimated using Cronbach alpha.
[a] Multiple-choice items and open-ended questions. [b]Five-point Likert-scale items.

4.3 Independent variables

The independent variables were SES and ethnicity, measured at both student level (level 1) and school level (level 2). The ethnicity of a student (E1) was based on the parents' place of birth as reported by the school's administrative office: Students with parents from European and other western countries received score 0 (ethnic majority), whereas other students received score 1 (ethnic minority). If the country of birth of one of the parents was unknown, the birthplace or SES of the other parent was used. If the countries of both parents differed, the mother's country of origin was selected. The ethnicity at school level (E2) was the proportion of ethnic minority students. School composition based on E2 ranged from 0 to 1 ($M = .22, SD = .41$).

SES at the student level was obtained from the parents' educational level. Three levels were distinguished: High SES (SH1) if the parents had a university degree or a degree in higher vocational education or university, medium SES (SM1) if the parents had a degree in intermediate vocational education or higher tracks of secondary school, and low SES otherwise. If for one parent information on the SES was missing, the SES of the other parent was selected. If the parents had different levels of SES, the mother's level of SES was selected. In the analyses, two dummy variables were used for SES at the student level with low SES being the reference level. SES at the school level was measured by the percentage of low-SES students. Table 2 (lower panel) shows a summary of the descriptive statistics of the dependent and independent variables.

4.4 Method of analysis

Regression analyses were used to describe the relation between SES and ethnicity (independent variables) and student proficiencies (dependent variables). School composition (SES and ethnicity measured at the school level) effects were estimated using random intercept models. For both static scores and gain scores, and each of the five dependent variables we considered three random-intercept models: no independent variables, independent variables at the student level, and independent variables at both student level and school level. Hence a total of 30 models were considered.

Let Y_{ij} be the value of a dependent variable for student i in school j, let μ be the fixed intercept, let U_{0j} be the random intercept for school j, en let e_{ij} be the error term. The three models were:

Null-model: Consists of an intercept but no predictors; that is,

$$Y_{ij} = \mu + U_{0j} + e_{ij}. \quad (1)$$

The null model was used as a benchmark to measure the effects of student-background characteristics and school composition.

Model 2: Consists of an intercept and SES and ethnicity at the student level as predictors; that is,

$$Y_{ij} = \mu + U_{0j} + b_1 \times SH1 + b_2 \times SM1 + b_3 \times E1 + e_{ij}. \quad (2)$$

Model 2 was used to explore the effect the student's own background has on their outcomes (both gains and static) but it does not adjust for the composition of the student's school.

Model 3: Consists of an intercept and SES and ethnicity as predictors both at the student level and at the school level; that is,

$$Y_{ij} = \mu + U_{0j} + b_1 \times SH1_{ij} + b_2 \times SM1_{ij} + b_3 \times E1_{ij} + b_4 \times SES2_j + b_6 \times E2_j + e_{ij}. \quad (3)$$

Model 3 was used to examine the amount of variance of the dependent variables that can be explained by school composition. Model 3 also provided the posterior residuals at the school level to measure the strength of the association between school composition and dependent variables. The models were analyzed using the software package MLwiN 2.25 (Rasbash, Steele, Browne, & Goldstein, 2015). In each of the three models, random effects were investigated using the intraclass correlation coefficient (ICC), which may be interpreted as the proportion of variance accounted for by the school (e.g., Snijders & Bosker, 2012). Let R_m^2 be the variance explained in Model m. Fixed effects were investigated using f^2 as follows: $f_{02}^2 = \frac{R_2^2}{1 - R_2^2}$ is the proportion of variance explained by the student effects compared to the proportion of outcome variance unexplained by Model 2, and $f_{23}^2 = \frac{R_3^2 - R_2^2}{1 - R_3^2}$ is the proportion of variance explained by the school effects, compared to the proportion of outcome variance unexplained by Model 3 (e.g., Aiken & West, 1991; Lorah, 2018), with f^2

$> .02$ for a small effect, $f^2 > .15$ for a medium effect, and $f^2 > .35$ for a large effect (Cohen, 1992; Lorah, 2018).

5. Results

Gain scores and static scores in well-being showed no significant correlation with SES and ethnicity at the student level. Therefore, well-being was excluded from the analyses. For each model, for both static and gain scores, and for each proficiency, Table 4 shows the intraclass correlation (ICC) for random effects, effect size R^2 for the fixed effects, and effect sizes f_{02}^2 and f_{23}^2. The complete output, including the parameter estimates of the fixed effects, is available in online Supplement S1.

Table 4
Random effects (ICC), Variance Explained (R^2), and Fixed Effects (f^2) for the Null Model, Model 2, and Model 3 on Each Proficiency

Proficiency	ICC			R^2		f_{02}^2	f_{23}^2
	Null	Model 2	Model 3	Model 2[a]	Model 3		
RG	.04	.03	.03	.04	.04	.04	.00
RS	.13	.08	.07	.13	.14	.15	.01
MG	.09	.08	.07	.02	.03	.02	.01
MS	.14	.10	.10	.10	.10	.11	.00
TG	.02	.03	.02	.00	.00	.00	.00
TS	.09	.09	.06	.00	.04	.00	.04
EG	.02	.02	.02	.01	.01	.01	.00
ES	.06	.05	.04	.03	.04	.04	.01

Note. RG = Reading comprehension gain, RS = Reading comprehension static, MG = Mathematic proficiency gain, MS = Mathematical proficiency static, TG = Task motivation gain, TS = Task motivation static, EG = Self-efficacy gain, ES = Self-efficacy static.
[a] $R_0^2 = 0$ by definition.

For static scores of mathematic proficiency and reading comprehension under the null model (Table 4, second and fourth row, first column), 14% and 13% of the variance, respectively, was accounted for by the school. For all other proficiencies, and under all other models, the percentage of variance accounted for by the school was smaller. For Model 2, R^2 values ranged from zero (task motivation) to .13 (reading comprehension, static scores), indicating that the effects of student background characteristics ethnicity and SES explained at most 13% of the variance in proficiency. Except for task motivation (static score), the R^2 values of Model 3 and Model 2 were similar, indicating that the additional explanatory value of school composition was limited. Effect sizes f^2 ranged from 0 (negligible effect, several cells in Table 4) to 0.15 (medium effect, reading comprehension static score). These effect sizes were used to answer our research questions.

5.1 Static versus gain scores.

As expected, the effect of student background was larger for static scores than for gain scores. The result was true for the cognitive proficiencies reading comprehension ($f_{02}^2 = .15$ v. $f_{02}^2 = .04$) and mathematical proficiency ($f_{02}^2 = .11$ v. $f_{02}^2 = .02$), and the non-cognitive proficiency self-efficacy ($f_{02}^2 = .04$ v. $f_{02}^2 = .01$). For task motivation, for both gain scores and static scores, student-level ethnicity and SES had no effect. Also, as expected, the additional effects of school composition were larger for static scores than for gain scores. However, except for task motivation ($f_{23}^2 = .04$ v. $f_{23}^2 = .00$), for both gain scores and static scores the effect sizes were negligible ($f_{23}^2 < .02$).

5.2 Cognitive versus non-cognitive proficiencies.

As expected, the effects of student background on cognitive proficiencies were larger than the effects of student background on non-cognitive proficiencies. Effect sizes (f^2) of student background ranged between .02 and .15 for cognitive proficiencies and between .00 and .04 for non-cognitive proficiencies. With the exception of the static score on task motivation ($f_{23}^2 = .04$; Table 4, eighth column), school composition only had negligible effects on both cognitive and non-cognitive proficiencies.

5.3 Student background versus school composition.

For the cognitive proficiencies and self-efficacy, the results confirmed our expectation that the effect of student background characteristics (Table 4, column 7) on these proficiencies is larger than the effect of school composition (Table 4, column 8). Contrary to our expectations, for the static score on task motivation the effect of student background ($f_{02}^2 = .00$) was smaller than the effect of school composition ($f_{23}^2 = .04$). Table 5 summarize the expected and observed effects.

Table 5

Expected and Observed Effects of Correction Method, Type of Proficiency, and Scoring Method on Cognitive and Non-Cognitive Proficiency Test Scores.

Scoring method	Correction method			
	Student background		School composition	
	Proficiency		Proficiency	
	Cognitive	Non-cognitive	Cognitive	Non-cognitive
Expected effects				
Static score	++++	+++	++	++
Gain score	+++	++	+	+
Observed effects				
Static score	+++	++/- [a]	++	++/- [a]
Gains score	++	++	+	+

Note. The number of +s indicate the relative strength of the positive effects. For example, +++ indicates a stronger positive effect than ++.
[a] With the exception of task motivation, for which the effect was negative.

6. Conclusion and discussion

We investigated the effects of student background characteristics and the combined effect of student background characteristics and school composition on cognitive and non-cognitive proficiencies. By and large, our findings matched our expectations, which were based on previous research (Table 5). We found that both student and school characteristics had larger effects on static scores than on gain scores, which indicates that static scores are more influenced by effects outside the realm of control of the school.

As expected, we found the effect of student background to be larger on cognitive proficiencies than on non-cognitive proficiencies. This is in line with results by Konu et al., (2002). Task motivation might by less influenced by external factors than the other proficiencies; no effect was found on both its static score and its gain score of student background and the effect of school composition was only found for static task motivation. But our results on this construct differ from other studies (e.g., Opdenakker & Van Damme, 2000; Ruzek, Domina, Conley, Duncan, & Karabenick, 2015) that did find a (small) correlation between student background, cognitive achievement and task motivation. Because the literature on school-level effects, like school composition based on SES on motivation, is scant (Urdan & Schoenfelder, 2006), making inferences is difficult.

Both gain scores and static scores in well-being showed no significant correlation with SES and ethnicity at the student level. Hence well-being was not analyzed further. Earlier research indicates that well-being is much more dependent of other factors and less on SES and ethnicity (Opdenakker & Van Damme, 2000; Ryan & Deci, 2000), which might explain why we did not find an effect either.

Estimating the added value of teachers in the context of accountability and evaluation requires a correction for both student-level and school-level characteristics. Therefore the preferred model is one that is least influenced by factors outside the sphere of influence of schools. This study adds to the ongoing debate about the use of VAMs using gain scores by showing that school composition does not significantly affect gain scores of cognitive proficiencies, whereas it does influence static and corrected static scores, indicating that gain scores might by less influenced by factors outside the sphere of influence of a school.

The merits of this study are the following. First, we added to the body of literature a strong argument to, when evaluating schools on the bases of student proficiencies, to use gain score VAM's, correcting for student background. Second, the availability of students' scores on both cognitive and non-cognitive proficiencies enabled us to investigate the differences between the two types of proficiencies. Third, our large data set met the advocated sample-size criteria of multilevel analyses, and enabled us to analyze the data properly.

This study also had some limitations. First, we investigated effects at the school level whereas the effects of class level may have been larger. For example, Schneeweis and Winter-Ebmer (2007) showed that the effect of class composition on student achievement is larger than the effect of school composition. Second, in this study, tests were administered in the third and sixth grade, whereas a completer picture of students' gains and the value-added by schools might have been obtained when the progress during the students' entire school careers had been monitored and measured. The same can be said about taking into account the changes in their socioeconomic background during a student's career and thus between two measurements of proficiency (e.g., Strand, 1997).

The results of this study support and add to the current body of literature on the use of different correction models when using student proficiencies in accountability. They offer a new argument for researchers and policy makers to use VAMs based on gain scores. The argument holds that gain scores are significantly less associated with student background and especially school composition than the static scores. The results also show that cognitive proficiencies are more affected by student background and school composition than non-cognitive proficiencies. Lastly the results indicate that school composition adds only little extra in terms of correcting on top of student background characteristics. We did not have data about the composition of the specific classes, the effects of class composition may be larger than school composition, this will need further research. More research is also needed to investigate the inverse outcomes found for task motivation, where school composition appeared to influence gain scores more than static scores.

References

Bankston, C., & Caldas, S. J. (1998). Family structure, schoolmates, and racial inequalities in school achievement. *Journal of Marriage and Family*, *60*(3), 715–723. https://doi.org/10.2307/353540

Bill & Melinda Gates Foundation. (2014). Ensuring fair and reliable measures of effective teaching. In *MET Project Research Paper*. https://doi.org/10.3886/ICPSR34771.v2

Bosker, R. J. (2012). De toegevoegde waarde van een school; begripsbepaling, meting en causale atributie [the value added of a school; Definition, measurement, and causal attribution]. In A. B. Dijkstra & F. B. G. Janssens (Eds.), *Om de kwaliteit van het onderwijs: Kwaliteitsbepaling en kwaliteitsbevordering* (pp. 93–104). Boom Lemma.

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, *53*, 371–399. https://doi.org/10.1146annurev.psch.53.100901.135233

Braun, H. I. (2005). Using Student Progress To Evaluate Teachers : A Primer on Value-Added Models. *Princeton NJ Educational Testing Service*, 20. http://www.etsliteracy.org/Media/Research/pdf/PICVAM.pdf

Caldas, S. J., & Bankston, C. (1998). The inequality of separation: Racial composition of schools and academic achievement. *Education Administration Quarterly*, *34*, 533–557. https://doi.org/10.1177/0013161X98034004005

Cito. (2015). *Wetenschappelijke verantwoording van de LVS-toetsen Begrijpend lezen tweede generatie [Scientific accountability for reading comprehension second generation]* (Issue november).

Coe, R., & Fitz-Gibbon, C. T. (1998). School effectiveness research: criticisms and recommendations. *Oxford Review of Education*, *24*, 421–438. http://www.jstor.org/stable/1050665

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155. https://psycnet.apa.org/fulltext/1992-37683-001.html

Driessen, G. (2002). School composition and achievement in primary education: a large-scale multilevel approach. *Studies in Educational Evaluation*, *28*, 347–368.

Driessen, G. (2007). *'Peer group' effecten op onderwijsprestaties [Peer group effets on educational achievement]. November.*

Driessen, G. (2009). *Prestaties , gedrag en houding van basisschoolleerlingen [Achievement, behaviour and attitude of primary school students].*

Driessen, G., Mulder, L., Ledoux, G., Roeleveld, J., & van der Veen, I. (2009). *Cohortonderzoek COOL5-18: Technisch rapport basisonderwijs, eerste meting 2007/08 [Cohort study COOL5-18: technical report primary education, first measurement]*. 156.

Driessen, G., Mulder, L., & Roeleveld, J. (2012). *Cohortonderzoek COOL5-18: Technisch rapport basisonderwijs, tweede meting 2010/11 [Cohort study COOL5-18: technical report primary education, second measurement].*

Dronkers, J., & Robert, P. (2008). School choice in the light of the effectiveness differences of various types of public and private schools in 19 OECD countries. *Journal of School Choice*, *2*(3), 260–301. https://doi.org/10.1080/15582150802371499

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, *82*(1), 405–432. https://doi.org/10.1111/j.1467-8624.2010.01564.x

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*(4), 341–349. https://doi.org/10.1037//1040-3590.8.4.341

Faubert, V. (2009). *School evaluation: Current practices in OECD countires an a literature review* (Issue 42). https://doi.org/10.1787/9789264117006-8-en

Feenstra, H., Kamphuis, F., Kleintjes, F., & Krom, R. (2010). *Wetenschappelijke verantwoording Begrijpend lezen voor groep 3 tot en met 6 [Scientific accountability reading comprehansion for grades 3 to 6]* (Issue november).

Gabriel, R., & Allington, R. (2012). The MET project: The wrong 45 million dollar question. *Educational Leadership*, *70*(3), 44–49.

Goldstein, H., & Cuttance, P. (1988). A note on national assessment and school comparisons. *Journal of Education Policy*, *3*(2), 197–202. http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/national-assessment-and-school-comparisons.pdf

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, *100*(2), 267–271. https://doi.org/10.1257/aer.100.2.267

Harker, R., & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement*, *15*(2), 177–199.

Harris, D. N. (2011a). Value-added measures and the future of educational accountability. *Science*, *333*, 826–827. https://doi.org/10.1126/science.1193793

Harris, D. N. (2011b). Value-added measures in education; what every educator needs to know. In *Science* (Vol. 333). Harvard Education Press. https://doi.org/10.1126/science.1193793

Hollenberg, J., van der Lubbe, M., & Sanders, P. (2011). *Toetsen op school primair onderwijs [Testing in school primary education]*.

Houtveen, A. A. M. A. M., van de Grift, W., & Creemers, B. P. M. P. M. (2004). Effective school improvement in mathematics. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice*, *15*(3–4), 337–367. https://doi.org/DOI: 10.1080/09243450512331383242

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 accountability of the tests LOVS Aritmathics-Matehmatics for grades 3 to 8]*.

Kelly, S., & Majerus, R. (2011). School-to-School variation in disciplined inquiry. *Urban Education*, *46*(6), 1553–1583. https://doi.org/10.1177/0042085911413151

Konu, A. I., Lintonen, T. P., & Autio, V. J. (2002). Evaluation of well-being in schools? A multilevel analysis of general subjective well-being. *School Effectiveness and School Improvement*, *13*(2), 187–200. https://doi.org/10.1076/sesi.13.2.187.3432

Koppen, J. K., Lunt, I., & Wulf, C. (2002). Education in Europe; Cultures, values, institutions in transition. *European Studies in Education*, *14*, 54–74.

Lorah, J. (2018). Effect size measures for multilevel models: definition, interpretation, and TIMSS example. *Large-Scale Assessments in Education*, *6*(1), 8. https://doi.org/10.1186/s40536-018-0061-2

Maas, C. J. M., & Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-1881.1.3.86

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*(2), 127–137. https://doi.org/10.1046/j.0039-0402.2003.00252.x

McConney, A., & Perry, L. B. (2010). Science and mathematics achievement in Australia: The role of school socioeconomic composition in educational equity and effectiveness. *International Journal of Science and Mathematics Education*, *8*(3), 429–452. https://doi.org/10.1007/s10763-010-9197-4

Midgley, C., Maehr, M. L. L., Hruda, L. Z. Z., Anderman, E., Anderman, L., Freeman, K. E. E., Gheen, M., Kaplan, A., Kumar, R., Middleton, M. J. J., Nelson, J., Roeser, R., & Urdan, T. (2000). *Manual for the Patterns of Adaptive Learning Scales (PALS)*.

OECD. (2008). *Measuring improvements in learning outcomes: Best practices to assess the* *value-added of schools*. https://doi.org/10.1787/9789264050259-en

OECD. (2015). *Skills for social progress: the power of social and emotional skills* (OECD Skills Studies). OECD Publishing. https://doi.org/10.1787/9789264226159-en

Opdenakker, M.-C., Van Damme, J., De Fraine, D. F., Van Landeghem, G., & Onghena, P. (2002). The effect of schools and classes on mathematics achievement. *School Effectiveness and School Improvement*, *13*, 399–427. https://doi.org/10.1076/sesi.13.4.399.10283

Opdenakker, M. C. C., & Van Damme, J. (2000). Effects of schools, teaching staff and classes on achievement and well-being in secondary education : similarities and differences between school outcomes. *School Effectiveness and School Improvement*, *11*(2), 165–196. https://doi.org/10.1076/0924-3453(200006)11:2;1-Q;FT165

Palardy, G. J. (2008). Differential school effects among low, middle, and high social class composition schools: a multiple group, multilevel latent growth curve analysis. *School Effectiveness and School Improvement*, *19*(1), 21–49. https://doi.org/10.1080/09243450801936845

Peetsma, T., Hascher, T., & van der Veen, I. (2005). Relations between adolescents' self-evaluations, time perspectives, motivation for school and their achievement in different countries and at different ages. *European Journal of Psychology of Education*, *20*(3), 209–225. https://doi.org/10.1007/BF03173553

Perry, L. B., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers College Record*, *112*(4), 1137–1162.

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

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of*

Educational and Behavioral Statistics, *20*(4), 307–335. https://doi.org/10.3102/10769986020004307

Rumberger, R. W., & Palardy, G. J. (2005). Does segregation still matter? The impact of student composition on academic achievement in high school. *Teachers College Record*, *107*(9), 1999–2045. https://doi.org/10.1111/j.1467-9620.2005.00583.x

Ruzek, E. A., Domina, T., Conley, A. M., Duncan, G. J., & Karabenick, S. A. (2015). Using value-added models to measure teacher effects on students' motivation and achievement. *The Journal of Early Adolescence*, *35*(5–6), 852–882. https://doi.org/10.1177/0272431614525260

Ryabov, I., & Van Hook, J. (2007). School segregation and academic achievement among Hispanic children. *Social Science Research*, *36*(2), 767–788. https://doi.org/10.1016/j.ssresearch.2006.04.002

Ryan, R., & Deci, E. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American Psychologist*, *55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Saunders, L. (1999). A brief history of educational "value added": How did we get to where we are? *School Effectiveness and School Improvement*, *10*(2), 233–256.

Scheerens, J. (2012). Wat is kwaliteit?[What is quality?]. In A. B. Dijkstra & F. J. G. Janssens (Eds.), *Om de kwaliteit van het onderwijs: Kwaliteitsbepaling en kwaliteitsbevordering* (pp. 43–76). Boom Lemma.

Seegers, G., van Putten, C. M., & de Brabander, C. J. (2002). Goal orientation, perceived task outcome and task demands in mathematics tasks: effects on students' attitude in actual task settings. *The British Journal of Educational Psychology*, *72*(Pt 3), 365–384. https://doi.org/10.1348/000709902320634366

Sloane, F. C., Oloff-Lewis, J., & Kim, S. H. (2013). Value-added models of teacher and school effectiveness in Ireland: Wise or otherwise? *Irish Educational Studies*, *32*(1), 37–67. https://doi.org/10.1080/03323315.2013.773233

Strand, S. (1997). Pupil progress during key stage 1: a value added analysis of school effects. *British Educational Research Journal*, *23*, 471–487. https://doi.org/10.1080/0141192970230406

Teddlie, C., & Reynolds, D. (2000). *International handbook of school effectiveness research*. Psychology Press.

Timmermans, A. C. (2012). *Value added in educational accountability: Possible, fair and useful?* http://dissertations.ub.rug.nl/FILES/faculties/gmw/2012/a.c.timmermans/Proefschrift.pdf

Tomesen, M., Engelen, R., & Hiddink, L. (2019). *Wetenschappelijke verantwoording Begrijpend lezen 3.0 voor groep 5 [Scientific accountability of the reading comprehension tests for grade 5]*.

Urdan, T., & Schoenfelder, E. (2006). Classroom effects on student motivation: Goal structures, social relationships, and competence beliefs. *Journal of School Psychology*, *44*(5), 331–349. https://doi.org/10.1016/j.jsp.2006.04.003

Van Der Slik, F. W. P., Driessen, G. W. J. M., & De Bot, K. L. J. (2006). Ethnic and socioeconomic class composition and language proficiency: A longitudinal multilevel examination in Dutch elementary schools. *European Sociological Review*, *22*(3), 293–308. https://doi.org/10.1093/esr/jci058

Verhaeghe, J. P., Damme, J. Van, & Knipprath, H. (2011). *Value added of primary schools with high proportions of minority students: a longitudinal study*. www.steunpuntloopbanen.be

White, & Whitehurst, G. J. (2013). Teacher value-added: Do we want a ten percent solution ?

The Brown Center Chalkboard, 1–8. www.brookings.edu/blogs/brown-center-chalkboard/posts/2013/04/24-merit-pay-whitehurst

Zvoch, K., & Stevens, J. J. (2008). Measuring and evaluating school performance: An investigation of status and growth-based achievement indicators. *Evaluation Review*, *32*(6), 569–595. https://doi.org/10.1177/0193841X08320398