WORKING PAPER
DANISH SCHOOL OF EDUCATION

BENT SORTKÆR AND DAVID REIMER

DISCIPLINARY CLIMATE AND STUDENT ACHIEVEMENT: EVIDENCE FROM SCHOOLS AND CLASSROOMS

AARHUS UNIVERSITET

ISBN: 978-87-7507-376-4 (Online)
DOI: 10.7146/aul.154.126

Bent Sortkær and David Reimer

Disciplinary Climate and Student Achievement: Evidence from Schools and Classrooms

Working paper
Danish School of Education, Aarhus University, 2016

Title:
Disciplinary Climate and Student Achievement: Evidence from Schools and Classrooms

Published by
Working paper
Danish School of Education, Aarhus University 2016

Authors:
Bent Sortkær* and David Reimer**

*Danish School of Education, Aarhus University, Niels Juels Gade 84, Aarhus N, Denmark.
Email: beso@edu.au.dk ; Telephone: +45-28802950

** Danish School of Education, Aarhus University, Niels Juels Gade 84, Aarhus N, Denmark.
Email: dare@edu.au.dk (corresponding author) ; Telephone: +45-87163935

Abstract:
Disciplinary climate has emerged as one of the single most important factors related to student achievement. Using data from the OECD Programme for International Student Assessment (PISA) 2003 for Canada, Denmark, Finland, Iceland, Latvia and Norway we find a significant and nontrivial association between the perceived disciplinary climate in the classroom and students' mathematics performance in Canada, Denmark and Norway. Furthermore we exploit country specific class-size rules in order to single out a subsample with classroom-level data (PISA is sampled by age and not by classes) and find that the estimates based on school-level data might underestimate the relationship between disciplinary climate and student achievement. Finally we find evidence for gender differences in the association between disciplinary climate and student achievement that can partly be explained by gender-specific perceptions of the classroom environment.

Keywords: Disciplinary Climate, Classroom Climate, Multilevel Analyses, PISA, Gender Differences

The research for this paper was partly funded by the Danish Council for Strategic Research (Research project CCIELO: Classroom composition, inclusion, exclusion and learning outcomes)

Indhold

Introduction

The disciplinary climate in schools or classrooms can be considered as one of the single most important factors related to student achievement (Hattie, 2009; Scheerens, 2005; Wang, Heartel, & Walberg, 1993). A multitude of studies report statistically significant as well as substantively important effects of classroom disciplinary climate on student achievement (Arum & Velez, 2012; Figlio, 2007; Frempong, Ma, & Mensah, 2012; Marks, 2010; Ning, Van Damme, Van Den Noortgate, Yang, & Gielen, 2015; Teodorović, 2011). Given the relevance of disciplinary climate for learning, some of the major international large scale assessment studies have incorporated measures for classroom climate in the student questionnaires. Recent reports published by the OECD (2010, 2013b) as well as Ning et al. (2015), all based on PISA 2009 data from 2009, report a strong association between disciplinary climate and student performance. In the latter paper, the authors find that 11% of the between-school differences in reading achievement over countries can be explained by the classroom disciplinary climate for the 52 countries in their sample.

While scholars agree on the importance of disciplinary climate for learning outcomes, no consensus has emerged regarding the analytic level at which to measure the disciplinary climate construct. Driven by theoretical considerations in some studies and from what seems to be the availability of data in other cases, both the school, the age of students, the class and the individual student level have been used to measure disciplinary climate. Overall, few studies, have tried to gauge the relative benefits or disadvantages associated with the use of the different levels of measurement (but see Lüdtke, Trautwein, Kunter, & Baumert, 2007). By exploiting class-size rules in a number of selected OECD countries we perform analyses with the grade as well as the class level as unit of measurement for disciplinary climate. These analyses might also shed light on the question whether and to what degree data from the PISA study can be used as a classroom-level dataset. Consequently, this paper aims to contribute to the debate concerning the choice of analytic level for measuring disciplinary climate.

Another issue overlooked in the literature on classroom disciplinary climate is the question regarding potential gender differences. While differences between boys and girls in educational attainment and other learning outcomes such as grades or test scores have caught the attention of researchers and policy makers (Diprete & Buchmann, 2013; Weaver-hightower, 2003), few studies, have explored to what extent

the classroom climate potentially contributes to this phenomenon. Most analyses about the relationship between disciplinary climate and student achievement typically include gender as a control variable only. As a result, a second goal of this paper is to examine whether there is a gender difference in the strength of the relationship between disciplinary climate and student achievement. Furthermore we explore whether possible gender differences can be attributed to gender specific perceptions of the classroom environment given that girls might view the classroom climate more positive than boys do (Goh & Fraser, 1998) or whether there is evidence for a gender specific effects of the same classroom climate.

Against this background we will provide multilevel estimates of the relationship between the disciplinary climate and student achievement in mathematics among 9th or 10th graders in six western countries using data from PISA 2003 and examine the role of level of analyses and gender differences. The remainder of the paper is organized as follows. Section two reviews the literature on the relationship between disciplinary climate and learning outcomes, focusing on measurement issues and gender disparities. Section three describes the data, methods and measures used in the analyses. Section four presents the results of our analyses. Finally, in section results are summarized and discussed.

Literature Review

Measuring disciplinary climate

Student self-reports are commonly used in the evaluation of various aspects of the classroom. They are both inexpensive and easy to collect and the student's view of the classroom environment presents a unique perspective (den Brok, Brekelmans, & Wubbels, 2004; Gentry, Gable, & Rizza, 2002; Peterson & Stevens, 1988). While much research has been done in testing the reliability and validity of postsecondary student reports (Greenwald, 1997; Marsh, 1987), the literature for the compulsory school level is sparse. Ebmeier, Jenkins & Crawford (1991) test the validity of student reports by comparing how students and teaching experts evaluate the same teachers. The study concludes that the students and the experts agree in 93 % of the cases (or in all but one case).

In a longitudinal study Peterson & Stevens (1988) test the reliability of compulsory school students self-reports and finds that the students are being both consistent and stable over a two year span of time. Furthermore the study concludes that the students were able to discriminate among different teachers. Also using data from the United States, Polikoff (2015) test the reliability of student (grade 4-8) response on instructional quality and find that student evaluations are stable but the stability is lower than in comparable studies among tertiary students.

Another way to look at the question of validity of student reports is to draw on the theoretical concept of *perspective-specific validities* (Kunter & Baumert, 2007). In an analysis of data from a German extension to the 2003 PISA study, in which 288 mathematics teachers and their students participated, they examined to what extend teachers and students agreed on various aspects of instruction. Whereas most research is based on the assumption that there is one underlying true construct that cannot be measured accurately by neither the student nor the teacher (e.g. Ebmeier et al., 1991; Olsen, 2003) the concept of perspective-specific validities helps to explain how students and teachers within the same class seemingly experiences the same reality in different ways. We presume that perspective specific validities are also relevant to explain inter-individual variation between students. Students' perception of the disciplinary climate within the same classrooms might very well differ due to a variety of personal characteristics such as for example academic aptitude or gender. Nevertheless, previous studies showed that students within the same classroom will agree to some extent about

the way they perceive the classroom climate (Ebmeier et al., 1991; Kunter & Baumert, 2007; Peterson & Stevens, 1988).

Disciplinary climate, student achievement and level of analyses

In contrast to questions regarding interrater reliability and the validity of student vs. teacher ratings, the question at which organizational level measurement of disciplinary climate and related concepts should ideally take place has received very little attention in the literature.

Using the *school level* as the unit of analysis, Lassen et al. (2006) and Luiselli et al. (2005) find that elementary and middle school students in the United States improve their reading and mathematics skills as school disciplinary climate improves following an intervention program. The size of the effect reported in these studies might be somewhat overestimated as the intervention also includes guidelines for improving instructions in math and language classes which in itself might lead to learning gains.

Building on data from PISA several studies use *student age* to construct a measure for disciplinary climate. Even though the tested students are asked how they experience the disciplinary climate in the classroom, the sampling design of the PISA survey precludes classroom level analyses as there is no information collected on which class students attend. The result is a range of analyses based on all 15-year old students from the sampled schools. In these analyses the implicit assumption is made that all students from one grade experience and contribute to the same class-level disciplinary climate (Frempong et al., 2012; OECD, 2005b; Olsen, 2003; Rangvid, 2003; Välijärvi, Kupari, & Linnakylä, 2007). Only Olsen (2003) addresses this issue by including the headmaster's view on the disciplinary climate in the school in his analysis and arguing for the presence of a school-specific disciplinary climate culture. All these studies find that a better disciplinary climate is positively related to student achievement.

We only identified two studies that measure disciplinary climate at the classroom level. Goh and Fraser (1998) analyze data from 10-11 years old students in Singapore and find that the classroom mean for one disciplinary climate measure (friction) has a significant relationship with mathematics achievement while three other classroom climate measures do not. Teodorović (2011) finds that an orderly climate in the classroom is positively related to both mathematics and Serbian language achievement in primary schools in Serbia.

While most studies do not theoretically reflect the choice of analytical level for the measurement of classroom climate, the few studies that discuss this issue all favor the use of the classroom level (Ammermueller & Pischke, 2009; Marsh et al., 2012; Teodorović, 2009; Willms, 2006). Given that learning still mostly takes place in class-

room settings, we also expect that using the class as level of measure will yield the most precise estimate for measures of the classroom environment.

Finally, Marks (2010) uses the *individual student* as level of measure to predict the performance of the students when trying to enter tertiary education in Australia. Drawing on longitudinal data from PISA 2003 he finds a significant relationship between the disciplinary climate as perceived by the students in 9th grade and how well they do in a cognitive test when trying to enter tertiary education. In this paper we attempt to use both the grade as well as the class level as unit of measurement for disciplinary climate in order to gauge the relative benefits of one over the other analytical level when drawing on PISA data.

Gender differences

Few studies have examined whether the relationship between disciplinary climate and learning is heterogeneous across different groups of students - e.g. boys and girls. The studies we surveyed looked at how differences in gender (Goh & Fraser, 1998; Kuperminc, Leadbeater, Emmons, & Blatt, 1997) and in race and gender (Koth, Bradshaw, & Leaf, 2008) produce different perceptions of the school climate. Most of these studies find that girls perceive the classroom climate more positively than boys do. However, none of these studies have examined if these differences in perception lead to a differentiated relationship of the disciplinary climate on academic performance which seems reasonable to expect from a social cognitive point of view (Koth et al., 2008). We cast some light on this issue by exploring whether boys and girls experience the disciplinary climate differently as well as by exploring whether the disciplinary climate has a differentiated effect on these two groups of students[1]. In one of the few studies related to this topic Legewie & DiPrete (2012) show that boys are more affected by favorable as opposed to unfavorable classroom SES composition than girls. They argue that an academically oriented environment in schools "suppresses boys' negative attitudes toward school, and facilitates academic competition as an aspect of masculine identity" (Legewie & DiPrete, 2012, p. 468). We expect similar gender difference for disciplinary climate, e.g. we hypothesize that boys are more susceptible to unfavorable learning environments than girls are.

[1] There exist some research into whether cognitive ability and gender explain the disruptive behavior of a student but the results are ambiguous. Nordahl and colleagues (2009) find no relationship between the cognitive ability and disruptive behavior whereas Kaplan and colleagues (2002) find that being male and having lower achievement was associated with disruptive behavior. Further, Arum and colleagues (2012) finds that schools with a greater concentration of boys experience more disciplinary problems.

Data, methods and measures

Data

We draw on the PISA study from the year 2003, which offers a unique insight into the relationship between disciplinary climate and student achievement among 15-year olds. Other large scale international assessment studies such as TIMMS (Trends in International Mathematics and Science Study) or PIRLS (Progress in International Reading Literacy Study) also include some measures for the disciplinary climate. However, the classroom climate measure in these studies focus more on bullying or experiences of victimization rather than indicators for the disciplinary climate that allow for teaching in an orderly atmosphere. Besides focusing on the mathematical ability of the students, PISA 2003 includes information on a range of student and school characteristics as well as how students perceive the disciplinary climate in the classroom in math lessons. This information is gathered from questionnaires completed by the students and the headmasters of the schools. These properties make PISA 2003 tailored for our analyses.[2]

PISA 2003 uses a two-stage sampling design where the first stage sampling units are schools with 15-year-old students enrolled and the second stage sampling units are 15-year-old students. Based on the PISA 2003 data we create two samples. First we generate our analytic or full sample based on multiple as well as single class schools. In addition to this full sample we create a comparable subsample containing single class schools only. The procedure and the selection criteria for creating the two samples are outlined in four steps below.

Out of the 41 countries participating in PISA 2003 we only select countries with a sufficient degree of heterogeneity among the students at every school and in every classroom in terms of their cognitive ability. If a country has a high level of structural differentiation among students the relationship between disciplinary climate and student achievement could be a result of the homogeneity of schools and classes. The result would be a spurious correlation between disciplinary climate as perceived by the students and student achievement (Kaplan, Gheen, & Midgley, 2002). To ensure a sufficient degree of student heterogeneity we only select countries that do not track students before the age of 16, have a low degree of grade repetition, have few students out of modal starting age and have a low proportion of schools that group student by abili-

[2] PISA has a three year cycle but the 2006 edition has no information on the disciplinary climate and in 2009 three of our selected countries were using an explicit stratification sampling procedure which reduces the heterogeneity of the data (OECD, 2012).

ty in all subjects (Brunello & Checchi, 2006; European Communities, 2003; OECD, 2010). In the second step we remove students not attending the modal grade level to avoid unnecessary noise in the measurement of disciplinary climate. In Denmark, for example, this implies the exclusion of all students not attending 9th grade. In a third step we exclude schools with less than ten students in the sample to reduce possible biases due to outliers and to ensure a reliable aggregated disciplinary climate construct (Kunter & Baumert, 2007).

Finally, to select schools with only one class in the relevant grade we use information from the school questionnaire on the total number of students enrolled at each school combined with information on the number of grades at the school. In order to identify single class schools, we look for schools with an average of no more than 25 students per grade. If this criterion is fulfilled, then students are most likely not split into two classes.[3] This last selection criterion is crucial for the forming of our single class school subsample and is based on a careful investigation of the school systems and class-size rules in each of the selected countries.[4] For our analyses we identified *Canada, Denmark, Finland, Iceland, Latvia and Norway* as the countries that fit our institutional requirements regarding tracking coupled with sufficient student heterogeneity across schools and that have a suitably large number of single class schools that warrant further analyses. While the selection of countries was mostly driven by statistical concerns, the resulting sample is relatively homogenous and represents highly developed Western nations with extensive education systems which reduce problems related to comparability of country- specific findings.[5]

After excluding cases with missing values and conducting the selection procedure described above our analytic sample is based on 37,156 observations (68% of the original sample) attending either 9th or 10th grade depending on the schooling system in the country whereas our single class subsample consists of 2,850 observations (5% of the original sample).

[3] According to our procedure a school with 225 students and 9 grades (225/9) will have an estimated average of 25 students per grade. Even though that this is no bulletproof procedure, we deem it very likely that most 15 year old students selected through this rule will attend the same class.

[4] After conducting an extensive literature and document search as well as email correspondence, we operate with the following class size rules in the year 2003: Canada; while there is variation between provinces we chose 30 as upper limit; Denmark; 28, Finland; no limit, Iceland; 32, Latvia; 34, and Norway; 30. More documentation on the country-specific class size regulations is available on request.

[5] According to the Human Development Index, Canada, Denmark, Finland, Iceland and Norway were among the top 14 most developed societies in 2003 (UNDP, 2003). Latvia, which can also be labeled as highly developed, was placed somewhat lower on the rankings (50).

Methods

We start out by using the grade level measure for the disciplinary climate drawing on the full sample. In subsequent analyses the smaller single class subsample is used. The assumption underlying the grade level analysis is that in a grade at any given school there is a distinct disciplinary climate culture indicating that the disciplinary climate varies between schools but very little within schools in a given grade whereas the second part of the analysis allows each class to have its own classroom climate culture. Although our single class school subsample is not perfectly representative (see appendix A1 for a sample comparison) it can be used as a useful point of comparison with the full (grade-based) sample.

We run multi-level models (Snijders & Bosker, 2012) separately for each country using the individual student as level one and the school as level two. We start by running a null model to estimate the within-, and between-school variance and then add covariates on both analytic levels:

$$\gamma_{ij} = \beta_0 + \beta_1 SC_j + \beta_2 ST_{ij} + \mu_j + \varepsilon_{ij}$$

where γ_{ij} is the math score of student i in school j, SC_j are school level characteristics including a measure of the disciplinary climate, ST_{ij} are student level characteristics including gender, μ_j are unobserved characteristics of school j, and ε_{ij} are unobserved characteristics of student i within school j. For our classroom level analysis SC_j will then present classroom level characteristics. Finally we add on a cross-level interaction between disciplinary climate and gender to examine a possible gender specific difference in the relationship between these two.

Measures

Our main dependent variable is the PISA math test score of the students. The scale is derived from five plausible values and the weighted average OECD mean of these are 500 with a standard deviation of 100 (OECD, 2005a).

The measure for the disciplinary climate is based on student responses on how often the following things happen in their mathematics lessons: “students don’t listen to what the teacher says”; “there is noise and disorder”; “the teacher has to wait a long time for students to quieten down”; “students cannot work well”; and, “students don’t start working for a long time after the lesson begins”. These answers are used to build a student level disciplinary climate index which is standardized and centered with a grand mean of zero and a variation of one. A higher number indicates a more positive

disciplinary climate. To construct the aggregated school level averages we use information from all students in the relevant grade who reported a value for the relevant variable in the data set, not just the students in our final sample. By using this procedure we lower the sampling error in our aggregated disciplinary climate measure that is due to non-respondents (Lüdtke, Marsh, Robitzsch, & Trautwein, 2011). This combined with PISA 2003's very high response rate[6] leads to a very low level of sampling error - especially in our single class sample in which all relevant student are included in the survey. This aggregated school level construct too has been standardized and centered with a grand mean of zero and a variation of one.

We control for student and family characteristics at the student level as well as classroom and school characteristics at the school level (Table 1).

The variables for gender, language spoken at home, country of birth and school type (public vs. private) are dummy coded, the variables urban ranging from 1 (rural) through 5 (a large city with >1,000,000 citizens) and percentage bilinguals ranging from 1 (<10% bilingual) through 4 (>40% bilingual students) are on ordinal scales whereas the variables measuring SES at the individual and at the school level are continuous. The SES is a composite measure based on the student response on parental educational level (coded as years of schooling according to ISCED classification), parental occupational status (based on ISEI classification) and number of home possession including books in the home (OECD, 2005a, p. 316). The student level SES measure has been standardized with a grand sample mean of zero and a variation of one. The variable meanSES is an aggregated school level average of the SES variable.

[6] Weighted student participation rate after replacement are: Canada 84%, Denmark 90%, Finland 93%, Iceland 85%, Latvia 94% and Norway 88% (OECD, 2005a, p. 173)

Table 1. Descriptive statistics

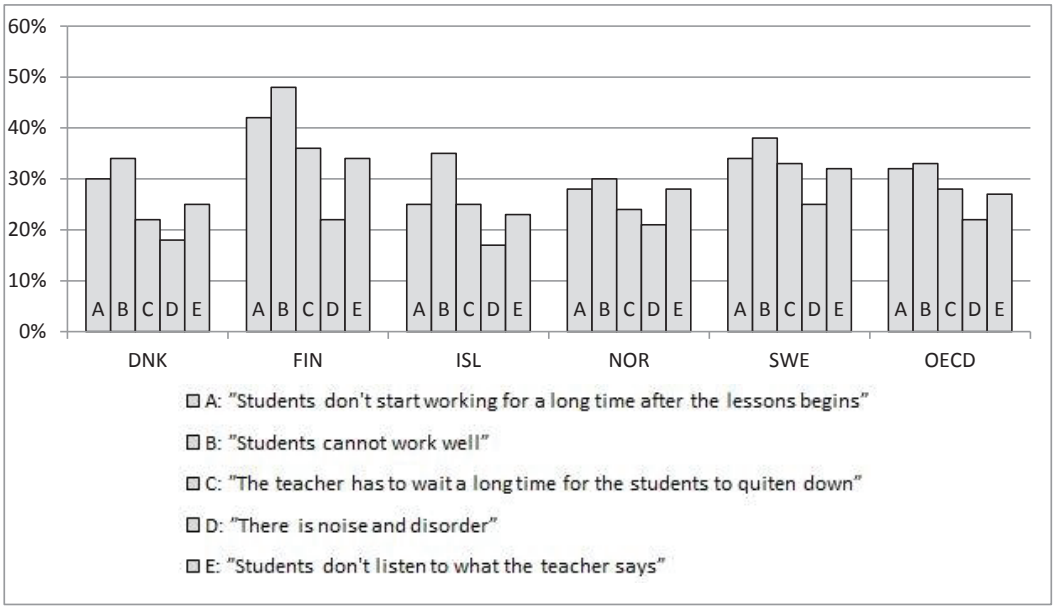
	Canada		Denmark		Finland		Iceland		Latvia		Norway	
Variable	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Dependent variable												
scoreMATH	535.01	78.96	522.61	82.58	550.83	76.15	517.99	85.46	496.95	77.89	500.03	86.69
Independent variables												
Student characteristics												
Female	.53	.50	.52	.50	.51	.50	.49	.50	.53	.50	.50	.50
SES	-.01	1.00	-.20	1.00	-.13	.99	.39	.97	-.23	.87	.26	.94
speak test-language	.94	.23	.97	.17	.99	.10	.98	.13	.99	.07	.96	.20
born test-country	.94	.23	.96	.21	.97	.16	.94	.24	.97	.18	.95	.21
School characteristics												
classdisc	.12	.94	-.14	.96	-.22	.92	-.33	.89	.81	1.11	-.56	.78
meanSES	-.01	.46	-.21	.45	-.13	.40	.39	.40	-.23	.37	.25	.36
Urban	2.53	1.25	2.34	1.02	2.62	.98	2.55	1.04	2.86	1.16	2.24	1.12
Public	.94	.24	.81	.39	.93	.25	1	0	1	0	.99	.10
percentage bilinguals	1.57	1.03	1.39	.77	1.22	.67	1.23	.77	1.62	1.07	1.27	.66
Number of students	19,541		3,272		4,865		2,657		3,199		3,622	
Number of schools	804		168		190		76		133		158	

Source: PISA 2003, Own calculations.

Results

Before turning to our multivariate analysis, we look at the prevalence of different aspects of disciplinary climate in mathematics lessons in the selected countries as reported by the students based on the variables used to construct the disciplinary climate index. There are considerable differences between the six selected countries and based on Figure 1 below it seems obvious that disciplinary climate – or the lack thereof – is an issue in all of the selected countries although not too different from the OECD average.

Figure 1. Percentage of the students who report that these incidents happen in most or every lesson



Using the individual students PISA math score as dependent variable in a multi-level (null model) we obtain the between and within school variation in math score from which we calculate the intraclass correlation (ICC(math)) displayed in Table 2. The intraclass correlation in math of the six countries is relatively low compared to other

Western countries indicating that the chosen countries have indeed very heterogeneous schools when it comes to mathematics achievement.[7]

Table 2. Within- and between school variance. Full Sample. PISA-score in math as dependent variable

	Canada	Denmark	Finland	Iceland	Latvia	Norway
τ^2 : between groups	1006.81	864.50	311.42	317.09	1168.89	497.61
σ^2 : within groups	5110.27	6025.35	5509.62	7026.50	4853.24	7010.99
Total	6117.08	6889.85	5821.04	7343.59	6022.13	7508.59
ICC(math)	.16	.13	.05	.04	.19	.07

Grade level analysis

Using the full sample and the grade as level of measure for the disciplinary climate we estimate a random intercept model with fixed effects. In Table 3 we see that there is a significant and nontrivial association between the disciplinary climate in the classroom and the students' mathematics performance in Canada, Denmark and Norway, while the parameter estimates in Finland, Iceland and Latvia are not statistical different from zero. A coefficient estimate of 12.67 in Canada, for example, translates to an improvement of 12.67 points in the PISA test for all students in the grade if the disciplinary climate in the classrooms improves by one standard deviation holding all other independent variables fixed. The estimate in Denmark and Norway is 9.13 and 9.08 respectively.

[7] The OECD intra-class average is .359 (OECD, 2005b)

Table 3. Multilevel regression models on student achievement in mathematics

	Canada			Denmark			Finland			Iceland			Latvia			Norway			Pooled sample		
	Null model	Full model	Int. model	Null model	Full model	Int. model	Null model	Full model	Int. model	Null model	Full model	Int. model	Null model	Full model	Int. model	Null model	Full model	Int. model	Null	Full	Interact.
Intercept	548.73***	545.35***	545.28***	522.18***	504.83***	504.83***	551.49***	512.57***	512.63***	516.47***	459.59***	460.92***	491.84***	541.93***	540.67***	499.15***	494.61***	495.54***	537.18***	521.41***	521.43***
	(1.96)	(9.43)	(9.42)	(2.73)	(11.53)	(11.55)	(1.80)	(21.55)	(21.53)	(2.85)	(16.49)	(16.65)	(3.64)	(19.80)	(19.83)	(2.28)	(10.07)	(10.14)	(1.45)	(6.66)	(6.64)
Student characteristics																					
Female		-16.43***	-16.23***		-17.33***	-17.58***		-10.54***	-10.98***		13.72**	11.39**		-10.18**	-4.32		-7.22*	-8.92*		-14.10***	-14.10***
		(1.88)	(1.93)		(2.76)	(2.77)		(1.92)	(1.93)		(4.57)	(4.32)		(3.02)	(3.63)		(3.06)	(3.84)		(1.23)	(1.23)
SES		20.11***	20.10***		26.43***	26.42***		25.09***	25.08***		23.45***	23.33***		19.37***	19.25***		34.09***	34.07***		23.19***	23.18***
		(.93)	(.94)		(1.49)	(1.49)		(1.18)	(1.12)		(1.63)	(1.62)		(1.76)	(1.73)		(1.59)	(1.59)		(.63)	(.63)
speak test-language		11.60**	11.60**		9.00	9.00		37.45*	37.60*		41.34*	41.47*		-2.14	-4.17		-.38	-.37		10.30**	10.29**
		(4.26)	(4.26)		(8.71)	(8.70)		(16.00)	(16.07)		(16.62)	(16.59)		(17.18)	(17.24)		(9.22)	(9.24)		(3.60)	(3.60)
born test-country		2.72	2.70		28.91***	28.95***		16.01*	16.01*		-.27	-.26		-13.37	-13.25		29.97***	30.02***		7.80**	7.79**
		(3.59)	(3.59)		(7.14)	(7.14)		(7.66)	(7.67)		(7.65)	(7.72)		(10.02)	(9.92)		(7.84)	(7.85)		(2.85)	(2.85)
School characteristics																					
classdisc		12.67***	13.45***		9.13***	10.06***		2.79†	3.52		3.45	7.21†		5.07	8.96*		9.08***	10.59**		10.97***	12.42***
		(1.58)	(1.87)		(1.78)	(2.51)		(1.65)	(2.29)		(2.44)	(3.73)		(3.32)	(4.13)		(2.30)	(3.07)		(1.14)	(1.31)
meanSES		18.81***	18.78***		24.71***	24.74***		7.98†	8.04†		8.42	8.12		41.69***	41.59***		8.79	8.90		11.50***	11.49***
		(4.09)	(4.09)		(4.55)	(4.56)		(4.84)	(4.85)		(7.13)	(7.16)		(9.08)	(9.08)		(6.23)	(6.23)		(2.90)	(2.91)
Urban		-1.05	-1.05		-1.82	-1.80		-4.24*	-4.25*		-.08	-.14		-2.07	-2.10		1.25	1.19		4.02***	4.01***
		(1.47)	(1.47)		(1.68)	(1.69)		(2.03)	(2.03)		(2.69)	(2.70)		(2.88)	(2.88)		(2.12)	(2.12)		(1.07)	(1.07)
Public		-5.04	-5.06		7.46	7.47		11.58	11.65		0	0		0	0		-21.80***	-21.84***		-6.32	-6.29
		(7.31)	(7.30)		(4.94)	(4.95)		(10.74)	(10.74)		(omitted)	(omitted)		(omitted)	(omitted)		(3.18)	(3.18)		(4.71)	(4.69)
percentage bilinguals		.05	.05		.30	.31		-1.79	-1.80		.97	1.10		-5.30*	-5.26*		-5.18†	-5.15†		-.88	-.87
		(1.66)	(1.66)		(1.98)	(1.98)		(1.99)	(1.99)		(1.77)	(1.81)		(2.37)	(2.36)		(2.81)	(2.81)		(1.39)	(1.39)
Cross level interaction																					
classdisc#girl			-1.44			-1.73			-1.43			-6.83			-7.31**			-2.98			-2.74*
			(1.75)			(2.98)			(2.09)			(4.78)			(2.68)			(4.05)			(1.10)
Number of students	19,541	-	-	3,272	-	-	4,865	-	-	2,657	-	-	3,199	-	-	3,622	-	-	37,156		
Number of schools	804	-	-	168	-	-	190	-	-	76	-	-	133	-	-	158	-	-	1,529		
Derived estimates																					
R^2	-	.14	.14	-	.20	.20	-	.12	.12	-	.09	.10	-	.13	.13		.17	.17		.14	.14
R^2 - student level	-	.08	.08	-	.12	.12	-	.11	.11	-	.07	.07	-	.06	.06		.14	.14		.09	.09
R^2 - school level	-	.45	.45	-	.77	.77	-	.33	.33	-	.55	.55	-	.42	.42		.58	.58		.34	.34

Note: Models estimated by maximum likelihood. Standard error in parentheses. The coefficient for public could not be estimated due to an insufficient number of private schools in the Icelandic and Latvian sample.
† p<0.10; * p<0.05; ** p<0.01; *** p<0.001 (two-tailed tests).

Estimates for the other independent variables in the model point in the expected direction, e.g. SES has a statistical significant and positive relationship with student achievement in all countries whereas not being born in the country of the test language or not speaking the test language at home is associated with lower mathematics scores in some of the selected countries. Boys seem to do better in mathematics than girls, with the exception of Iceland where girls outperform boys.
Looking at the school level variables Canadian, Danish and Latvian students do better in mathematics if they attend a school with a high mean SES. The variable urban shows a negative trend (only not in Norway) but is only statistical significant in Finland. Attending a public school in Norway is negatively associated with student achievement whereas a high percentage of bilingual students in the school have a negative relationship with student achievement in Latvia.

Class-level analysis

This second part will be using the subsample with single class schools based on the assumption that using the class as level of analysis will lead to a more precise estimate of the relationship between disciplinary climate and student achievement. As mentioned before, our sample is now reduced to 2,850 students.

The quality of this subsample becomes apparent when looking at the intra-class correlation (ICC) using the student level disciplinary climate variable as dependent variable in a multilevel null model (Table 4). As expected, the agreement among schoolmates as to their rating of the disciplinary climate in the classroom increases substantially in all countries but Denmark compared to the estimates based on the full sample. In our view this indicates that there are differences in disciplinary climate between classes within grades and could be an argument for using class-level data in classroom environment analyses.

Table 4. Within- and between school variance. Perceived disciplinary climate as dependent variable

	Canada	Denmark	Finland	Iceland	Latvia	Norway
Full sample						
τ^2 : between groups	.0885	.1132	.1076	.1404	.1652	.0722
σ^2 : within groups	.9371	.8173	.8689	.7731	.7334	.7975
total	1.0256	.9306	.9765	.9135	.8986	.8698
ICC	.09	.12	.11	.15	.18	.08
Single-class sample						
τ^2 : between groups	.1728	.0589	.1059	.2793	.1925	.1356
σ^2 : within groups	.7689	.8721	.6512	.6297	.5510	.6357
total	.9417	.9310	.7571	.9090	.7435	.7713
ICC	.18	.06	.14	.31	.26	.18

Looking at the intra-class correlation using the PISA score in mathematics as dependent variable in a multi-level null model the subsample from Canada seems to suffer from a growing homogeneity in the schools as to student achievement illustrated by a growth in the intra-class correlation from 0.16 in the full sample to 0.35 in the subsample (see appendix, Table A2). This could signify that the single class sample from Canada fail to live up to being heterogeneous at the school level meaning one should interpret the estimates of the Canadian subsample with caution. The five other countries sustain a heterogeneous distribution of students regarding their cognitive ability.

The result of re-estimating the random intercept model from part one of the analyses using the subsample as input is reported in Table 5. The relationship between disciplinary climate and PISA score is now statistical significant in Canada, Iceland, Latvia and Norway despite the smaller sample. Different from the full sample analysis is that the disciplinary climate coefficient estimate in Denmark no longer is statistical significant whereas the opposite is true for Iceland and Latvia. Comparing the magnitude of the estimated coefficients with the ones from the 'full model' (included in Table 5) the trend is a substantial strengthened relationship between disciplinary climate and student achievement with only Finland showing a smaller coefficient estimate. These results indicate that either (a) using the class as unit of analysis gives a more reliable measure of the disciplinary climate in the classroom and therefore a more precise estimate of the relationship between disciplinary climate and student achievement, or that (b) the change in coefficient size is due to systematic differences between the two samples.

Table 5. Multilevel regression models on student achievement in mathematics. Single-class sample

	Canada			Denmark			Finland			Iceland			Latvia			Norway		
	Null model	Full model	Single	Null model	Full model	Single	Null model	Full model	Single	Null model	Full model	Single	Null model	Full model	Single	Null model	Full model	Single
Intercept	518.33***	545.35***	531.93***	511.48***	504.83***	531.66***	526.78***	512.57***	483.83***	507.85***	459.59***	446.03***	475.62***	541.93***	636.23***	501.65***	494.61***	476.01***
	(10.32)	(9.43)	(25.37)	(9.97)	(11.53)	(46.83)	(7.08)	(21.55)	(39.90)	(5.21)	(16.49)	(28.56)	(9.44)	(19.80)	(40.81)	(7.82)	(10.07)	(37.33)
Student characteristics																		
Female		-16.43***	1.36		-17.33***	-31.34**		-10.54***	-2.80		13.72**	5.90		-10.18**	-17.60†		-7.22*	-.11
		(1.88)	(5.82)		(2.76)	(9.85)		(1.92)	(5.41)		(4.57)	(8.26)		(3.02)	(10.28)		(3.06)	(11.06)
SES		20.11***	17.40***		26.43***	23.00***		25.09***	24.09***		23.45***	24.65***		19.37***	24.45***		34.09***	36.66***
		(.93)	(3.01)		(1.49)	(4.38)		(1.18)	(6.52)		(1.63)	(3.82)		(1.76)	(3.33)		(1.59)	(5.56)
speak test-language		11.60**	16.77		9.00	-5.59		37.45*	0.00		41.34*	63.12*		-2.14	-55.62***		-.38	34.29
		(4.26)	(12.69)		(8.71)	(33.50)		(16.00)	(omitted)		(16.62)	(28.72)		(17.18)	(7.23)		(9.22)	(37.46)
born test-country		2.72	26.77		28.91***	24.13		16.01*	2.83		-.27	6.04		-13.37	-75.91†		29.97***	13.10
		(3.59)	(16.61)		(7.14)	(20.30)		(7.66)	(16.69)		(7.65)	(18.41)		(10.02)	(41.51)		(7.84)	(29.74)
School characteristics																		
classdisc		12.67***	17.49***		9.13***	15.40		2.79†	2.47		3.45	5.54*		5.07	12.35*		9.08***	20.72***
		(1.58)	(4.55)		(1.78)	(12.30)		(1.65)	(4.66)		(2.44)	(2.82)		(3.32)	(6.24)		(2.30)	(4.70)
meanSES		18.81***	88.61***		24.71***	27.91		7.98†	.42		8.42	18.06		41.69***	3.99		8.79	27.61
		(4.09)	(18.73)		(4.55)	(31.80)		(4.84)	(14.86)		(7.13)	(25.09)		(9.08)	(33.30)		(6.23)	(23.40)
Urban		-1.05	-36.57***		-1.82	-4.17		-4.24*	3.81		-.08	-9.28		-2.07	-20.01		1.25	1.02
		(1.47)	(7.55)		(1.68)	(6.82)		(2.03)	(10.41)		(2.69)	(12.19)		(2.88)	(13.26)		(2.12)	(7.33)
Public		-5.04	31.48†		7.46	7.75		11.58	65.30*		0.00	0.00		0.00	0		-21.80***	
		(7.31)	(18.46)		(4.94)	(19.30)		(10.74)	(28.61)		(omitted)	(omitted)		(omitted)	(omitted)		(3.18)	(omitted)
percentage bilinguals		.05	-4.28		.30	4.38		-1.79	-2.48		.97	4.66†		-5.30*	8.46		-5.18†	-15.51*
		(1.66)	(3.63)		(1.98)	(46.83)		(1.99)	(5.07)		(1.77)	(2.81)		(2.37)	(15.25)		(2.81)	(7.80)
Number of students	1,213	-	-	287	-	-	332	-	-	412	-	-	287	-	-	319	-	-
Number of schools	79	-	-	18	-	-	17	-	-	27	-	-	20	-	-	17	-	-
Derived estimates																		
R^2	-	.14	.27	-	.20	.17	-	.12	.11	-	.09	.11	-	.13	.15		.17	.20
R^2 - student level	-	.08	.05	-	.12	.12	-	.11	.08	-	.07	.10	-	.06	.12		.14	.14
R^2 - school level	-	.45	.67	-	.77	.48	-	.33	1.00	-	.55	.31	-	.42	.34		.58	.98

Note: Models estimated by maximum likelihood. Standard error in parentheses. The coefficient for public could not be estimated due to an insufficient number of private schools in the Icelandic, Latvian and Norwegian sample.

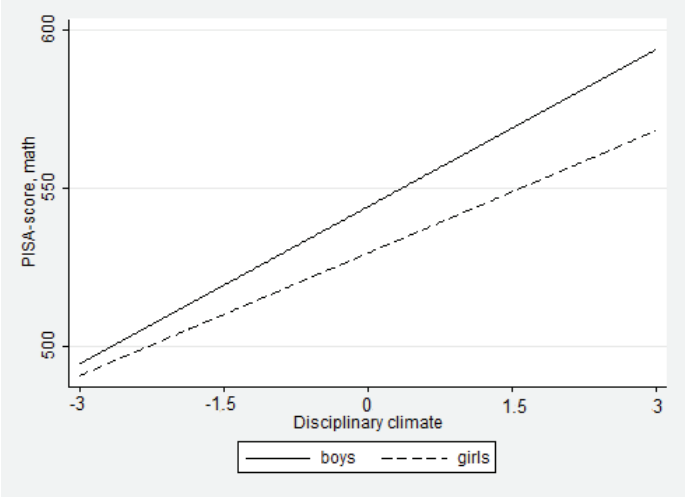
As a consequence of the selection procedure, the schools in our subsample are mostly smaller schools with a somewhat different SES composition (see sample comparison in Table A1). This could lead to a potential bias in the estimates if the mechanisms behind the relationship between disciplinary climate and PISA score are of different nature in smaller schools. However, the ICC coefficients show that in all countries but Canada, classrooms remain sufficiently heterogeneous in terms of student performance in the PISA tests which leads us to conclude that the single school sample can be useful to gain further analytical insight into classroom level phenomena.

Gender analysis

The final part of the paper will again draw on the full (school-level) sample. In this model we add a cross-level interaction between gender and disciplinary climate to the equation. As outlined before, we expect boys to be more susceptible to a negative disciplinary climate. The coefficient estimates for the interaction partially confirm this hypothesis. The sign of the interaction is negative in all six countries but with the exception of Latvia the coefficients are not estimated precisely enough to reach statistical significance.

Given that the interaction between gender and disciplinary climate seems to be similar across countries we run another model with a pooled sample (Table 3) which indicates that across all countries, boys are significantly more affected by the disciplinary climate than girls are, e.g. for each unit on the disciplinary climate index, boys' mathematics score falls by an additional 2.74 points (0.03 *SD*) compared to girls. The differential slope for girls and boys is illustrated in figure 2.

Figure 2. The relationship between disciplinary climate and math-score by gender. Pooled sample



To explore whether this gender differentiated relationship between disciplinary climate and student achievement is due to a gender specific difference in perception, we use the student level disciplinary climate index as the dependent variable in a multilevel random intercept model and test whether girls experiences the disciplinary climate in the classroom different than boys do (Table 6).

Table 6. Multilevel regression on disciplinary climate as perceived by the individual student as dependent variable

	Canada	Denmark	Finland	Iceland	Latvia	Norway
Female	.22***	.08*	.05	.10**	.16***	.04
	(.02)	(.03)	(.03)	(.04)	(.04)	(.03)
Number of students	19,541	3,272	4,865	2,657	3,199	3,622
Number of schools	804	168	190	76	133	158

Note: Models estimated by maximum likelihood. Standard error in parentheses. In all six countries we controlled for scoreMATH, SES, spoken language, country of birth, urban, school type and percentage bilingual.

† p<0.10; * p<0.05; ** p<0.01; *** p<0.001 (two-tailed tests).

Assuming boys and girls are equally distributed among schools[8], the coefficient estimated and p-values indicate that girls are experiencing a more positive disciplinary climate than boys in Canada, Denmark, Iceland and Latvia in spite of attending the same school and potentially the same classroom. This gender difference in perception is in line with the findings of Goh & Fraser (1998), Koth et al. (2008) and Kuperminc et al. (1997) and can serve as an explanation of the underlying mechanism causing boys to be more sensitive to the disciplinary climate than the girls. Part of the gender difference between boys and girls might thus be attributable to gender differences in the way the classroom-environment is perceived. To address this issue we compute a new, gender-specific, disciplinary climate index. Instead of using the school mean as the aggregate measure for disciplinary climate, we compute gender-specific indices, e.g. one average score for boys and one score for girls which is only based on the boys' and girls' responses, respectively.[9]

The results with the gender-specific disciplinary climate score are quite telling (Table 7). First the coefficient estimate for disciplinary climate and mathematic achievement is now significant in all countries but Latvia. Further the size of the estimates has 'evened out' across countries. County differences in the association between disciplinary climate and mathematics seem at least partly attributable to gender differences in the way the classroom environment is perceived. Finally the interaction between gender and disciplinary climate in the pooled model is also significant when using the gender-specific indices. We tentatively conclude from the results that the interaction between gender and disciplinary climate can be decomposed into two parts. Part of the association is due to gender differences in perception: girls view the same classroom climate more positive than boys do – and part seems to be a real "effect" in the sense that girls seem to be less affected by a negative classroom climate than boys. [10]

[8] There are only very few exceptions from this assumption in our data.

[9] It should be noted that strictly speaking the new gender specific index is not a level-2 but a level 1 variable since the climate score is not the same for all level 1 units (students) belonging to the respective level 2 units (schools).

[10] Using the gender-specific index for the disciplinary climate in this class-level model as we did in the grade level model above have two effects: The association between disciplinary climate and mathematics achievement in
Latvia is no longer significant. Also similar to the analyses based on the full sample, the size of the association becomes more similar across countries (results available on request).

Table 7. Multilevel regression on student achievement in mathematics using gender specific disciplinary climate measures

	Canada	Denmark	Finland	Iceland	Latvia	Norway	Pooled
Model 1:							
Gender specific disciplinary							
climate measure	7.26***	7.24***	3.75**	5.12*	4.65†	10.06***	6.57***
Model 2:	(1.41)	(1.73)	(1.35)	(2.60)	(2.66)	(1.95)	(.97)
Gender specific disciplinary							
climate measure - interaction	-1.82	-1.39	-3.35	-10.06†	-8.39**	-5.61	-3.26***
	(1.85)	(3.20)	(2.12)	(5.62)	(2.95)	(3.79)	(1.21)
Number of students	19,541	3,272	4,865	2,657	3,199	3,622	37,156
Number of schools	804	168	190	76	133	158	1,529

Note: Models estimated by maximum likelihood. Standard error in parentheses. In both models all six countries we controlled for scoreMATH, SES, spoken language, country of birth, urban, school type and percentage bilingual.

† $p<0.10$; * $p<0.05$; ** $p<0.01$; *** $p<0.001$ (two-tailed tests).

Discussion and Conclusion

Using data from PISA 2003 we find a statistical significant and non-trivial relationship between disciplinary climate and mathematics test achievement in Canada, Denmark and Norway among 15-year students attending the same grade indicating that in these countries a better disciplinary climate is associated with a better performance in the PISA math assessment. Furthermore, using a gender-specific index for the measurement of disciplinary climate, we find coefficient estimates reach statistical significance in all studied countries but Latvia.

By re-estimating the analysis using a subsample of single class schools we find a strengthened relationship between disciplinary climate and student achievement. This might indicate that the parameter from the full sample is possibly downwardly biased due to the use of the grade as level of measurement instead of the class. Although we cannot completely rule out that this finding is related to differences between small vs. all type of schools, we find it unlikely that this is the underlying cause for the diverse results. Consequently, we conclude that classroom level sampling, which is the case in some other large scale international assessment studies such as PIRLS and TIMSS, has clear advantages when exploring the consequences of classroom climate.

We also examined whether there is a gender difference in the magnitude of the relationship between disciplinary climate and student achievement. We find that in Latvia boys are more negatively affected by the disciplinary climate than girls are. The trend in the other countries is the same but the gender difference in these countries does not reach statistical significance. A pooled sample reveals a statistical significant gender difference across all six countries. Serving as a partial explanation of this phenomenon, we find that girls perceive the disciplinary climate more positively than boys in Canada, Denmark, Iceland and Latvia.

Such as is the case for most analyses based on large scale assessment data such as the PISA study, there are some concerns related to the validity of our presented findings. Due to the cross-sectional nature of the analyzed data, reciprocity between disciplinary climate and learning is quite possibie (Frenzel, Pekrun, & Goetz, 2007; Zimmermann, Schütte, Taskinen, & Köller, 2013). Furthermore, omitted variables might mediate the relationship between classroom climate and learning. Disciplinary climate could also be the result of other underlying explanatory factors such as disorganized teaching or the absence of classroom management. But even if that were to be

the case, we consider the classroom climate to be an important mediator variable for students learning outcomes.

While these methodological concerns are not new, the presented analyses raise important questions in relation to school environment research. In light of the gender differences we discovered, it seems to be advisable to account for relevant group differences in perception and effects of the classroom environment in future research. In addition to gender, class, race and other stratification variables might lead students to experience the classroom in different ways and thus mediate the effects of classroom environment. Furthermore, the issue regarding level of measurement is important and should be paid more attention to in future research. The sheer availability of data at the grade level, such is the case in PISA, should not drive researchers' decisions about level of measurement. While the OECD is aware of these limitations (OECD, 2013a) – the PISA consortium should be encouraged to move into the direction of class-level sampling. This would also enable to verify to what extend the magnitude of cross country differences in the association between disciplinary climate and learning (e.g. Ning et al. 2015) remains stable once the analyses are based on classroom- rather than school-level measurements.

References

Ammermueller, A., & Pischke, J.-S. (2009). Peer effects in european primary schools - Evidence from the Progress in International Reading Literacy Study. *Journal of Labor Economics*, *27*(3), 315–348.

Arum, R., & Velez, M. (2012). *Improving Learning Environments - school discipline and student achievement in comparative perspective*. Stanford: Stanford University Press.

Brunello, G., & Checchi, D. (2006). *Does School Tracking Affect Equality of Opportunity ? New International Evidence* (No. 2348).

Den Brok, P., Brekelmans, M., & Wubbels, T. (2004). Interpersonal Teacher Behaviour and Student Outcomes. *School Effectiveness and School Improvement*, *15*(3-4), 407–442.

Diprete, T. A., & Buchmann, C. (2013). *The rise of women: The growing gender gap in education and what it means for American schools*. New York: Russell Sage Foundation.

Ebmeier, H., Jenkins, R., & Crawford, G. (1991). The predictive validity of student evaluations in the identification of meritorious teachers. *Journal of Personnel Evaluation in Education*, *4*(4), 341–357.

European Communities. (2003). *Education across Europe 2003*.

Figlio, D. (2007). Boys named Sue: Disruptive children and their peers. *Education Finance and Policy*, *2*(4), 376–394.

Frempong, G., Ma, X., & Mensah, J. (2012). Access to Postsecondary Education: Can Schools Compensate for Socioeconomic Disadvantage? *Higher Education*, *63*(1), 19–32.

Frenzel, A. C., Pekrun, R., & Goetz, T. (2007). Perceived learning environment and students' emotional experiences: A multilevel analysis of mathematics classrooms. *Learning and Instruction*, *17*(5), 478–493.

Gentry, M., Gable, R. K., & Rizza, M. G. (2002). Students' perceptions of classroom activities: Are there grade-level and gender differences? *Journal of Educational Psychology*, *94*(3), 539–544.

Goh, S., & Fraser, B. (1998). Teacher interpersonal behaviour, classroom environment and student outcomes in primary mathematics in Singapore. *Learning Environments Research*, (January), 199–229.

Greenwald, a G. (1997). Validity concerns and usefulness of student ratings of instruction. *The American Psychologist*, *52*(11), 1182–6.

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

Kaplan, A., Gheen, M., & Midgley, C. (2002). Classroom goal structure and student disruptive behaviour. *The British Journal of Educational Psychology*, *72*(Pt 2), 191–211.

Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2008). A multilevel study of predictors of student perceptions of school climate: The effect of classroom-level factors. *Journal of Educational Psychology*, *100*(1), 96–104.

Kunter, M., & Baumert, J. (2007). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, *9*(3), 231–251.

Kuperminc, G. P., Leadbeater, B. J., Emmons, C., & Blatt, S. J. (1997). Perceived School Climate and Difficulties in the Social Adjustment of Middle School Students. *Applied Developmental Science*, *1*(2), 76–88.

Lassen, S., Steele, M., & Sailor, W. (2006). The relationship of schoolâ€◎wide Positive Behavior Support to academic achievement in an urban middle school. *Psychology in the Schools*, *43*(6).

Legewie, J., & DiPrete, T. A. (2012). School Context and the Gender Gap in Educational Achievement. *American Sociological Review*, *77*(3), 463–485.

Luiselli, J. K., Putnam, R. F., Handler, M. W., & Feinberg, A. B. (2005). Whole-school positive behaviour support: effects on student discipline problems and academic performance. *Educational Psychology*, *25*(2-3), 183–198.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2× 2 Taxonomy of Multilevel Latent Contextual Models: Accuracy-Bias Trade-Offs in Full and Partial Error Correction Models. *Psychological …*, *16*(4), 444–467.

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2007). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research*, *9*(3), 215–230.

Marks, G. N. (2010). What aspects of schooling are important? School effects on tertiary entrance performance. *School Effectiveness and School Improvement*, *21*(3), 267–287.

Marsh, H. W. (1987). Students' evaluations of University teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, *11*(3), 253–388.

Marsh, H. W., Lüdtke, O., Nagenast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, *47*(2), 106–124.

Ning, B., Van Damme, J., Van Den Noortgate, W., Yang, X., & Gielen, S. (2015). The influence of classroom disciplinary climate of schools on reading achievement: a cross-country comparative study. *School Effectiveness and School Improvement*, (August), 1–26.

Nordahl, T., Mausethagen, S., & Kostøl, A. (2009). *Skoler med liten og stor forekomst av atferdsproblemer. En kvantitativ og kvalitativ analyse av forskjeller og likheter mellom skolene*. Elverum.

OECD. (2005a). *PISA 2003 Technical Report*. OECD Publishing.

OECD. (2005b). *School Factors related to Quality and Equity - results from PISA2000*.

OECD. (2010). *PISA 2009 Results: What Makes a School Successful? - Resources, Policies and Practices (Volume IV)*.

OECD. (2012). *PISA 2009 Technical Report*. OECD Publishing.

OECD. (2013a). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. *OECD Publishing*.

OECD. (2013b). *PISA in focus* (Vol. 09).

Olsen, R. V. (2003). Student and teacher behaviour. In S. Lie, P. Linnakylä, & A. Roe (Eds.), *Norther Lights on PISA - unity and diversity in the Nordic countries in PISA 2000* (pp. 113–122). Oslo.

Peterson, K. D., & Stevens, D. (1988). Student reports for school teacher evaluation. *Journal of Personnel Evaluation in Education*, *2*(1), 19–31.

Polikoff, M. S. (2015). The Stability of Observational and Student Survey Measures of Teaching Effectiveness. *American Journal of Education*, *121*.

Rangvid, B. S. (2003). Educational Peer Effects Quantile Regression Evidence from Denmark with PISA2000 data. *Do Schools Matter?*, (45).

Scheerens, J. (2005). Review of school and instructional effectiveness research.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (second edi). London etc.: Sage Publishers.

Teodorović, J. (2009). Educational effectiveness: Key findings. *Zbornik Instituta Za Pedagoska Istrazivanja*, *41*(2), 297–314.

Teodorović, J. (2011). Classroom and school factors related to student achievement: what works for students? *School Effectiveness and School Improvement*, *22*(2), 215–236.

UNDP. (2003). *Human Development Report 2003*. New York: Oxford University Press.

Välijärvi, J., Kupari, P., & Linnakylä, P. (2007). *The Finnish success in Pisa-and some reasons behind it: Pisa 2003. 2*.

Wang, M. C., Heartel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research*, *63*(3), 249–294.

Weaver-hightower, M. (2003). The " Boy Turn " in Research on Gender and Education. *Review of Educational Research*, *73*(4), 471–498.

Willms, J. D. (2006). *Learning Divides : Ten Policy Questions About the Performance and Equity of Schools and Schooling Systems* (No. 5). Montreal.

Zimmermann, F., Schütte, K., Taskinen, P., & Köller, O. (2013). Reciprocal Effects Between Adolescent Externalizing Problems and Measures of Achievement. *Journal of Educational Psychology*, (April), 1–16.