I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11268

For Better or Worse? The Effects of Physical Education on Child Development

Michael C. Knaus
Michael Lechner
Anne K. Reimers

JANUARY 2018

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11268

For Better or Worse? The Effects of Physical Education on Child Development

Michael C. Knaus
SEW, University of St. Gallen and IZA

Michael Lechner
SEW, University of St. Gallen, CEPR, PSI, CESifo, IAB and IZA

Anne K. Reimers
Chemnitz University of Technology

JANUARY 2018

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

For Better or Worse? The Effects of Physical Education on Child Development*

This study analyses the effects of regular physical education at school on cognitive skills, non-cognitive skills, motor skills, physical activity, and health. It is based on a very informative data set, the German Motorik-Modul, and identifies the effect by using variation in the required numbers of physical education lessons across and within German federal states. The results show improvements in cognitive skills. Boys' non-cognitive skills are adversely affected driven by increased peer relation problems. For girls, the results suggest improvements in motor skills and increased extra-curricular physical activities. Generally, we find no statistically significant effects on health parameters.

JEL Classification: I26, Z28, I12

Keywords: physical education, cognitive skills, non-cognitive skills, motor skills, physical activity, health

Corresponding author:
Michael Knaus
Swiss Institute for Empirical Economic Research (SEW)
University of St. Gallen
Varnbüelstrasse 14
9000 St. Gallen
Switzerland

E-mail: Michael.Knaus@unisg.ch

* We thank Beatrix Eugster, Jeff Smith and Carina Steckenleiter as well as participants at the annual conference of the European Association of Labour Economists in St. Gallen in 2017, the Essen Health Conference 2017, as well as seminars at the University of St. Gallen and in St. Anton for helpful comments and suggestions. The usual disclaimer applies.

1 Introduction

Almost every healthy student participates in compulsory physical education (PE) during her school days (UNESCO, 2014). Therefore, increasing the amount of compulsory PE seems to be a unique measure for policymakers to counteract physical inactivity and the resulting health problems of children. PE presents the only direct channel that influences physical activity for all students. In contrast, indirect channels like subsidies for sports clubs or investments into sports infrastructure target mostly students who are already physically active.

Thus, it is not surprising that politicians and health organisations, among others, frequently promote increases in PE. The US Surgeon General recommends time in PE of 150-225 min per week (Benjamin, 2010), while the average time of PE in the US is less than 90 min/week (Cawley, Frisvold, & Meyerhoefer, 2013).[1] In Europe, several countries discuss minimum PE levels of one PE lesson per school day, which would lead to a considerable increase compared to current levels. For example, Austria introduced daily PE lessons for all-day schools in 2015 and plans to extend it for all schools.[2] However, the empirical evidence about the effects of PE at school is scarce and inconclusive. This is unfortunate because increasing time in PE requires substantial investments in new facilities and teachers as well as rearrangements of the curricula. Furthermore, there may be implicit costs in terms of children's human capital, as the additional time in PE has to come either from reducing hours of other subjects or a reduction in the children's leisure time. Consequently, policymakers, parents, and children need reliable evidence regarding the effects of different numbers of PE lessons to be able to judge whether the potential benefits of a future policy change outweigh its costs. This paper provides some new information in this respect.

[1] Other US policy advisors ask for the same range (Institute of Medicine of the National Academies, 2013).

[2] See BGBl § 6 Abs. 4a.

The UNESCO (2014) analyses PE curricula worldwide. They find that most PE curricula intend to provide beneficial effects along five domains: (i) support cognitive skill development, (ii) foster personal and social development (non-cognitive skills), (iii) improve basic motor skills to enable participation in active society, (iv) encourage physical activity in and out of school, and (v) improve health. We are not aware of any study that investigates the effects of PE on all five domains. Of course, a large body of evidence documents the (short-term) effectiveness of school-based interventions on specific domains.[3] However, these kinds of interventions are usually not permanent and taught by specially trained staff. Thus, it is not clear whether their results carry over to standard PE taught by regular teachers in regular schools.

The identification of the effects of regular PE is complicated by potential selection into schools providing more or less PE. Parents and children might have preferences for more or less PE and choose schools accordingly. Further, the amount of PE could vary with the quality of schools. On the one hand, high quality schools could offer less PE and devote more time to academic subjects. On the other hand, high quality schools could provide more PE because they have a better infrastructure. Controlling for all these factors would be challenging and requires very detailed information about parents and schools.

We are aware of three studies that address the selection into PE by using instrumental variables. They all evaluate the effects of regular PE in the US (Cawley et al., 2013; Cawley, Meyerhoefer, & Newhouse, 2007; Dills, Morgan, & Rotthoff, 2011). These studies use variation in PE requirements across US states as instrumental variable for the actual amount of PE for students. Cawley et al. (2007) find that additional time in PE increases the weekly

[3] Such interventions are found to increase physical activity and reduce sedentary behavior (De Meester, van Lenthe, Spittaels, Lien, & De Bourdeaudhuij, 2009; Demetriou & Höner, 2012; Hynynen et al., 2016; Kriemler et al., 2011), increase health-related fitness knowledge (Demetriou, Sudeck, Thiel, & Höner, 2015), and improve health related outcomes (Quitério, 2012), but usually have no influence on BMI (Harris, Kuramoto, Schulzer, & Retallack, 2009; Lavelle, Mackay, & Pell, 2012).

activity level of students but has no effect on the body-mass-index (BMI) and the probability to be overweight. Dills et al. (2011) focus on academic achievements and find no statistically significant effects of increased PE on average test scores. Cawley et al. (2013) find that more PE decreases the prevalence of overweight and obesity for boys. However, all three studies have problems with the power of the instrument. This could explain the mostly statistically insignificant estimates. A different approach is taken by Sabia, Nguyen and Rosenberg (2016). They exploit PE requirement changes in six US states in a difference-in-differences setting to investigate effects on body weight and physical activity. These reforms led to changes in PE activity of less than 10 minutes per week. Therefore, it is not surprising that they could not document any statistically significant effects on body weight and only minor increases in moderate activity for boys.

Our study contributes to the very limited literature about the effects of regular PE in various ways. (i) Our unique dataset enables a comprehensive analysis of all five domains of intended PE effects. (ii) We use a new identification strategy, by exploiting differences in PE requirements across *and* within German federal states (*Länder)* to identify the causal effects of PE. It turns out that these differences provide a powerful instrument for PE. (iii) We estimate the effects using a semi-parametric instrumental variable (IV) estimator to avoid unnecessary functional form assumptions in the estimation. (iv) We document the robustness of our findings by providing a variety of sensitivity checks regarding the assumptions and implementation underlying our identification and estimation strategy.

Our results show substantial increases in cognitive skills, measured as school grades, but adverse effects on non-cognitive skills, measured as increasing behavioural problems. The adverse effects are observed only for boys, while girls benefit even in terms of lower emotional symptoms. This suggests gender differences in the effectiveness of PE. In addition, we find improved motor skills and increased extra-curricular physical activities for girls.

Effects on motor skills and extra-curricular physical activities seem to be much smaller, if not absent, for boys. Regardless of gender, we find no statistically significant effects on any health parameter.

The paper is structured as follows. The next section describes the institutional setting generating the exogenous variation that we exploit. Section 3 describes the data. Section 4 explains the empirical strategy. Section 5 shows descriptive statistics of the relevant variables. Section 6 presents the main results, some heterogeneity analysis, and discusses the sensitivity of the results. Section 7 discusses the results in light of the existing literature and offers potential explanations for the findings. Section 8 concludes. Further background material is provided in several appendices.

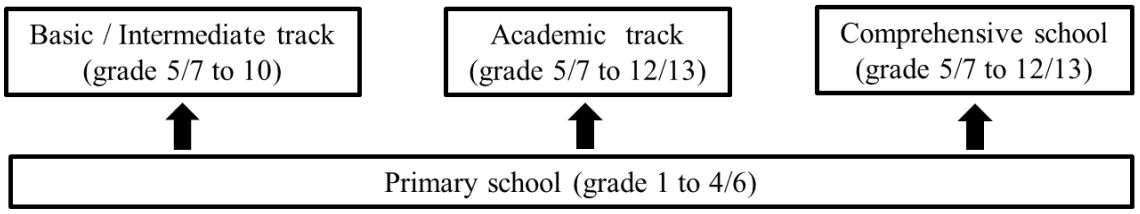
2 Institutional setting

The 16 German states enjoy high autonomy in specifying the details of the school system. We exploit the variation in the number of the PE lessons that is induced by the different states' regulations. Before going into detail, it is helpful to clarify the main features of the German school system. Figure 2.1 provides a stylised description. Children in Germany start their school career usually at the age of six in primary school (after non-compulsory kindergarten). After four to six years, school education continues in different secondary school tracks. For our analysis, we distinguish between three secondary school tracks: a basic / intermediate track, an academic track, and comprehensive schools. Some states split the first track further in a basic (*Hauptschule*) and intermediate track (*Realschule*). However, we cannot disentangle these two in our data and treat them as one track.[4] Both tracks last until grades 9 or 10. The academic

[4] The requirements are identical for basic and intermediate track in the states considered in our analysis. Thus, this shortcoming of the data does not influence our results.

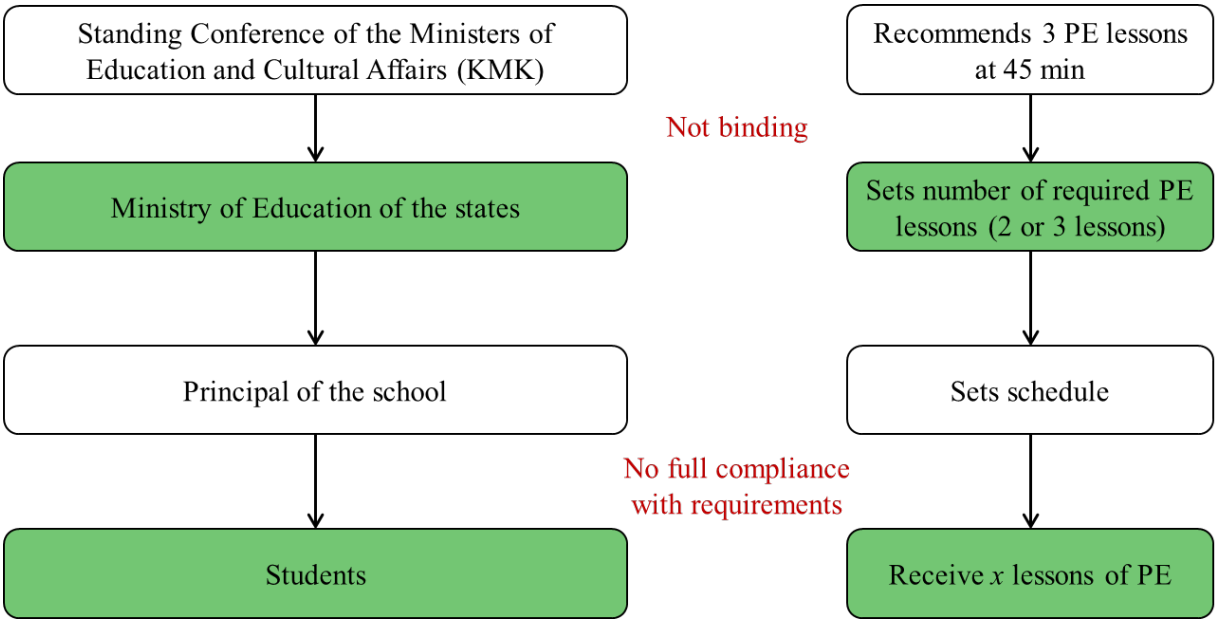
track (*Gymnasium*) lasts until grade 12 or 13. Additionally, comprehensive schools (*Gesamtschule*) combine the different tracks under one roof.

Figure 2.1: Stylised German school system



The details of the school system are determined within the states. However, the "Standing Conference of the Ministers of Education and Cultural Affairs" (KMK) formulates guidelines to foster comparable developments of the school system in all 16 states. These guidelines concern also curricula and thus the number of lessons of PE. Figure 2.2 illustrates the decision process and how these regulations actually influence the realised number of PE lessons. Generally, the KMK recommends three lessons of PE at 45 minutes per week. However, this recommendation is not binding. Binding curricula are formulated by the Ministries of Education of the states. These require either a minimum of two or three lessons of PE depending on state, school type, and class level. Coding of the specific requirements is provided in Appendix A. Principals of schools do not have to comply with the number of PE lessons required in the state-specific curricula. There are different possibilities of non-compliance going in different directions. Either schools have a sports profile and offer more PE than required, or shortages in facilities or staff prevent the realization of the required lessons of PE. The latter case prevails for states with three required lessons (Brettschneider, 2005). The described sequence of decisions determines the actual number of PE lessons that students experience.

Figure 2.2: Determination of PE lessons in Germany

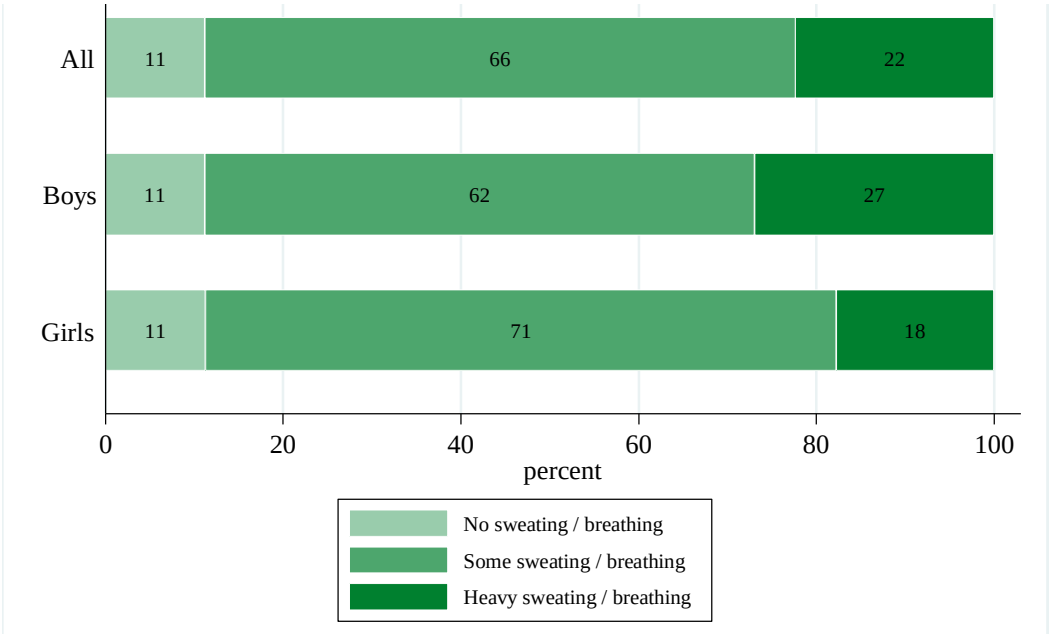


Below we analyse the effects of one additional PE lesson. Thus, a further description of German PE lessons is helpful to clarify what an additional PE lesson means for students in terms of additional activity, characteristics of a PE lesson, and potentially crowded-out education time in other subjects.

The activity survey, which is described in the next section in more detail, provides some information to describe how students perceive their PE lessons. The questionnaire in both waves asks about the physical intensity of PE. Figure 2.3 shows that the majority of students report only moderate activity during PE lessons with some sweating and breathing. This is in line with the observation in a validation study for the MoMo that reports rather moderate intensity of physical activity in regular PE (Jekauc, Wagner, Kahlert, & Woll, 2013), especially when compared to the intensity in club sports, which is substantially higher.[5]

[5] Similar patterns are observed for different studies in the UK and Denmark (see, e.g., Fairclough & Stratton, 2005; Møller et al., 2014).

Figure 2.3: Levels of activity in PE classes



Notes: Bars show the fraction of students reporting different categories of perceived intensity of physical activity on PE classes according the activity surveys in the Baseline and Wave 1 of the MoMo Study.

Table 2.1: Student's perception of PE lessons

Student's perception of PE lessons				
PE lessons are …		… neither … nor		PE lessons are …
… not important to me	8%	17%	75%	… important to me
… boring	7%	41%	52%	… varied
… not exhausting	17%	57%	26%	… exhausting
… easy	29%	62%	9%	… difficult
… chaotic	10%	39%	51%	… organised
… not movement-intensive	7%	31%	62%	… movement-intensive
… unstructured	8%	40%	52%	… structured

Notes: The questions are asked on a scale from 1-7. 1-2 are assigned to the left characteristics, 3-5 to neither nor, and 6-7 to the right characteristics. Based on 2,217 observations in Wave 1.

The activity survey of the second wave of the Motorik-Modul (MoMo) includes additionally detailed questions about the PE lessons. The students are asked how they perceive different characteristics of PE. The results in Table 2.1 show that the majority perceives PE as important for themselves, varied, neither exhausting nor not exhausting, neither easy nor difficult, organised, movement-intensive, and structured. The perceptions do not differ for students with two and three hour requirements. This indicates that students receive similar PE

lessons regardless of the required hours and that our effects are driven by additional PE lessons and not by different PE lessons.

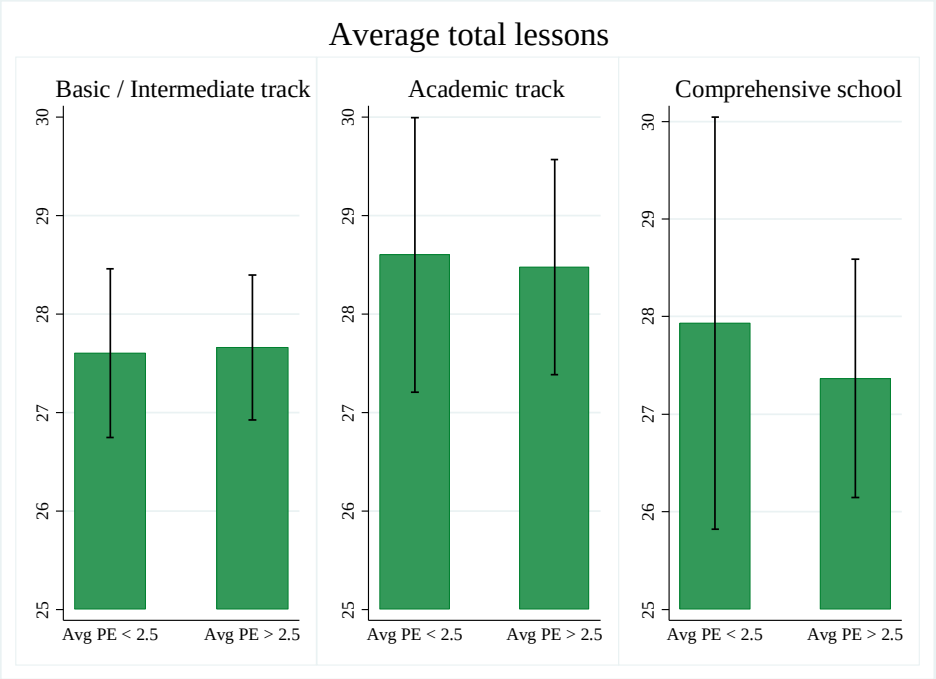
Finally, we address the question if more PE lessons mean longer total instruction time or crowding out of instruction time in other subjects. Unfortunately, the data provide no further information about the schedule of students besides compulsory and voluntary PE. This means that this question cannot be answered using the MoMo data. However, we collected curricula for all states and all tracks from the respective legal texts.[6] We extracted the required lessons for German, math, foreign languages, religion, music, arts, natural sciences, social sciences, and electives. We face the problem that curricula often state a cumulative number of lessons for several school years and it is impossible to assign an exact number to each class level. We deal with this by calculating the average number of lessons that students should attend in the school years 1 to 10 for each subject and track. We define groups of high and low PE states according to their average PE lessons being above 2.5 or below 2.5, respectively.[7] Figure 2.4 compares the average total amount of weekly lessons between high and low PE states. We find no evidence that more PE lessons result in longer total instruction time. The mean total lessons are very similar for the basic / intermediate and the academic track. In comprehensive schools, total lessons are on average even slightly shorter in high PE states.[8]

[6] In case of changes in the legal text during our sampling period, we use the status of 2012.

[7] If average PE requirements differ within states across tracks, we use the weighted average with weights according to the number of students observed in the respective tracks. This results in a difference of expected average PE requirements between high and low PE states of 0.7, which shows that dichotomizing expected average PE requirements at 2.5 discriminates well between high and low PE states.

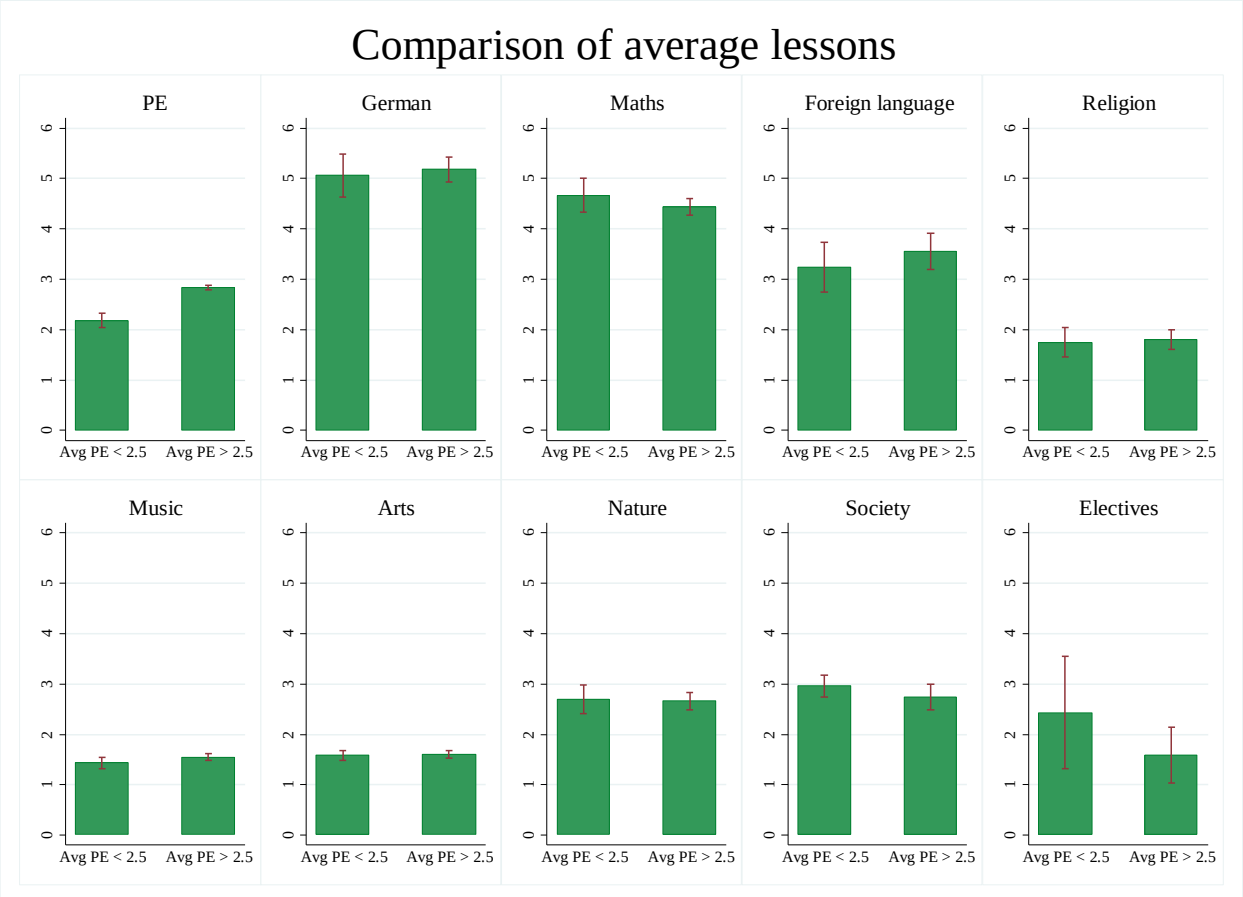
[8] This is in contrast to Cawley et al. (2013) who find for the US that an additional minute of PE increases the total length of school on average by 1.6 minutes.

Figure 2.4: Average number of total lessons



Notes: Bars show the average required number of total lessons over the first 10 school years averaged over states with average PE requirements below and above 2.5.

Figure 2.5: Comparison of average lessons



Notes: Bars show the average required number of subject lessons over the first 10 school years averaged over states with average PE requirements below and above 2.5.

This means that, at least on average, more PE lessons must crowd out some instruction time in other subjects. Figure 2.5 shows that the additional PE lessons seem to crowd out mostly elective courses. These tend to be a 3rd foreign language, or other elective specialisations in natural and social sciences. German and math, which are subject to our empirical analysis below, show remarkably similar number of lessons between high and low PE states.[9]

Voluntary PE lessons are not included in our measure (and are only used as an outcome later on).[10] Students are asked about them separately and it seems that students understood this distinction. We check this by comparing reported actual PE lessons for students with and without voluntary PE lessons in the same state, school type, and class level. Those attending voluntary classes report no systematically higher regular PE lessons.

3 Data

The data for the analysis stem from the Motorik-Modul (MoMo) (Wagner et al., 2014). The MoMo Study is a submodule of the longitudinal German Health Interview and Examination Survey for Children and Adolescents (KiGGS) (Kurth et al., 2008). While the KiGGS Study provides high quality health data and several measures for cognitive and non-cognitive skills, it lacks detailed information about physical activity and motor skills. However, this information is available in the MoMo Study, which is conducted for a subsample of the KiGGS participants. Questionnaires for both studies are answered by parents and children together (4-10 year old), or by the participants themselves (older than 10 years). The Baseline was conducted between 2003 and 2006 and the so-called Wave 1 from 2009 to 2012. The sampling procedure and data

[9] The graphical findings are confirmed in Table A.2.1 in Appendix A.2 by applying simple regressions and hold also after controlling for school type dummies.

[10] For results on the effects of participation in voluntary PE lessons see, e.g., Lunn and Kelly (2015).

preparation is described in detail in Appendix B. We work with 5,423 observations stemming from 4,698 individuals.

The extensive activity survey of the MoMo Study includes questions about habitual physical activity at school (Jekauc et al., 2013; Schmidt, Will, Henn, Reimers, & Woll, 2016). We use the question about the number of PE lessons to define our activity variable.[11] It is important to note that this question aims only at compulsory PE lessons and not at voluntary additional sports classes, which are asked about separately.[12]

The activity survey entails in addition the outcome variables that are used to capture the effects of PE on extracurricular physical activity. We observe three activity indices reflecting the habitual physical activity in club sports, leisure sports, and the sum of both indices labelled as extracurricular physical activity. Additionally, we observe if children participate in club sports at all, how many days per week they engage in moderate to vigorous physical activity for at least 60 minutes excluding PE, and if they comply with the WHO guidelines of *daily* 60 minutes of moderate to vigorous physical activity (WHO, 2010).

The MoMo Study includes additional measures for the other four outcome categories of interest. Cognitive skills are measured by German and math grades as well as the average of both. Non-cognitive skills are measured by means of the Strengths and Difficulties Questionnaire (SDQ) (Goodman, 1997). Motor skills are objectively measured using a battery of tests, which are applied in the MoMo Study (Woll, Kurth, Opper, Worth, & Bös, 2011).[13] The tests include assessments of strength, coordination, and stretchability. Finally, health

[11] The exact wording in the German survey is *"An wie vielen Tagen pro Woche hast du Sportunterricht in der Schule?"* (translation: *"How many days per week do you practice physical education at school?"*). Followed by the question that we use to construct our activity variable *"Wie viele Unterrichtsstunden (à 45 min) pro Woche sind das in der Regel zusammen?"* (translation: *"In total, how many lessons per week (at 45 minutes) are these in general?"*).

[12] Spengler, Mess, and Woll (2016) provide a detailed analysis of physical education and extracurricular sports activities measured in the MoMo Study.

[13] The test instructions in German are provided in Schmidt et al. (2016).

parameters are measured as subjective health (1-5), Body-Mass-Index (BMI), and resting heart rate.

The MoMo-data also provide rich socio-economic information about the students like household income, parent's education, parent's physical activity, household composition, nationality, birth weight, year of birth, degree of urbanisation, and educational spending per student at the state level.[14]

The requirements of PE lessons for each student are obtained from the statistical office of the KMK and double-checked with the corresponding legal texts on the state level.[15] The MoMo-data provide information about school type, class level, and state for each child. This enables us to merge the respective PE requirements to the students in our sample.

4 Empirical strategy

4.1 Identification

We are interested in the causal effect of PE lessons on a variety of outcomes. To this end, we exploit differences in required lessons of PE across states as an instrumental variable (IV) for the actual number of PE lessons experienced by students. Imbens and Angrist (1994) show that a valid instrument identifies the so-called local average treatment effect (LATE) in settings with a binary instrument and a binary treatment. Our application comes with a binary instrument because PE requirements are either two or three. However, the treatment variable of interest - number of PE lessons - ranges from zero to eight in our data and is thus discrete with bounded support. Frölich (2007) shows that a valid instrument in this setting identifies a weighted LATE.

[14] Information about spending per student is obtained from the Federal Statistical Office in Germany (https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/BildungKulturFinanzen/AusgabenSchueler.html).

[15] If both sources contained conflicting information, we relied on the legal text.

This weighted LATE represents in our application the average effect of an additional PE lesson for those students who actually receive more PE lessons because they live in states with higher PE requirements.[16]

This identification strategy requires three main assumptions to hold. First, *relevance* states that we observe at least some local compliance with required PE lessons. This is tested empirically below and turns out not to be problematic in this application. Second, the assumption of *monotonicity* rules out that students receive fewer PE because more lessons are required. In this particular environment, it does not seem plausible that school principals would schedule, e.g., three PE lessons if the curriculum requires two but two PE lessons if the curriculum requires three. Thus, monotonicity is a plausible assumption in this setting. The third assumption concerns the *exogeneity* of the instrument with respect to the considered outcome variables. Exogeneity means in our case that the different requirements must affect the outcome only through changes in the actual PE lessons and have no direct effect on the outcomes of interest. This identifying assumption is untestable and its plausibility must be thoroughly investigated.

We check the spatial distribution of average required PE lessons in the federal states in Figure A.3.1 of Appendix A.3. This reveals no obvious spatial clustering of high and low PE states. However, a closer look at the patterns in Tables A.1.1 and A.1.2 in Appendix A reveals that exogeneity may not hold unconditionally. The probability of three PE lessons depends on the grade of the students. Especially younger students have more often a requirement of three lessons. Additionally, the instrument varies across school types. Therefore, we control for class level and school type in our analysis because these factors affect outcomes of interest as well.

[16] Those students are weighted by their compliance intensity (Frölich, 2007). This means that students who receive one lesson more because three are required receive a weight of one in the weighted LATE, whereas students who receive two additional lessons receive a weight of two and so on. In this application, most students receive a weight of one because they are shifted from two to three lessons.

We investigate further whether students with more required PE lessons differ systematically in other observed characteristics that could also affect our outcome variables of interest. Such variables must also enter as control variables in our analysis to rule out that these differences invalidates the exogeneity assumption. We consider different socio-demographic, regional, and state characteristics to check whether their means differ by PE requirement and if they are significantly associated with a three hours PE requirement indicator in a logit regression. The results in Appendix C suggest that there is no selection into higher requirements with regard to household income, household composition, physically active parents, birthweight, and gender. However, we observe that higher PE requirements correlate significantly with higher education of parents, foreign status, year of birth, living in East Germany, urbanisation, and education expenditure per student. Consequently, we control for these differences in socio-economic, regional, and state characteristics in the analysis to establish exogeneity of the instrument at least conditionally.

Even after controlling for these observed factors, *policy endogeneity* may be a threat to our identification strategy. For example, benevolent policymakers in states with a relatively inactive youth might increase compulsory PE in school. We address this concern by comparing children of high and low PE states *before* they enter school. Fortunately, our dataset provides also information about 4 and 5 year old children who should not yet be affected by any PE requirements at school. To assess policy endogeneity concerns, we assign pre-school children to high PE states if the expected average PE lessons are higher than 2.5 throughout their school career and to low PE states if not. Most outcome measures are also available for pre-school children, with the obvious exceptions being grades and school-based activity. Appendix D provides the results of a simple unconditional mean comparison and a conditional mean comparison controlling for the characteristics by inverse probability tilting, which is described below. We find four significant unconditional differences at the 10 %-level and three significant differences at the 10 %-level for the conditional differences. This is about what we would expect

to occur by chance with the 49 tested outcomes. Therefore, policy endogeneity seems to be negligible for our application.[17]

Further, note that the number of required PE lessons are very stable over time. Table A.1.2 shows that we observe two changes in requirements over the nine years covered by our data. This might be surprising as the academic tracks underwent reforms in most states, decreasing the years in the academic track from nine to eight. This led to big changes in the curricula in general but left the required PE lessons mainly unchanged. In most states, they are already constant for decades. This strengthens the point that PE requirements are not endogenous in the sense that they are used as active policy measures to counteract specific developments in the states.

4.2 Estimation

The previous subsection shows systematic differences between students with high and low PE requirements. This motivates the need to control for a variety of characteristics in order to justify the exogeneity of the instrument at least conditional on observed characteristics. Additionally, we want to control for these characteristics in a flexible manner and avoid imposing unnecessary linearity conditions by applying two-stage least squares, for example. Therefore, we follow Frölich (2007) and estimate the weighted LATE (γ_w) as the ratio of the average treatment effect (ATE) of the instrument (Z) on the outcome of interest (Y) and the ATE of the instrument (Z) on the non-binary regressor (D):

$$\gamma_w = \frac{Z \rightarrow Y\ (ATE)}{Z \rightarrow D\ (ATE)}$$

[17] One sensitivity analysis additionally controls for the pre-school difference in the outcome variables. It shows that the detected significant differences we detect do not drive our results.

An estimator for the two ATE's entering this ratio should successfully balance the distribution of the considered control variables between the subsample with three and the subsample with two required PE lessons. We estimate the ATEs using inverse probability tilting (IPT) introduced by Graham, Pinto and Egel (2012). IPT is a variant of inverse probability weighting (IPW) that estimates the propensity score such that the means of the control variables in the subsamples with three and two required PE lessons are perfectly balanced.[18]

The control variables enter the propensity score in the following way: Class level dummies, school type dummies, gender dummy, eight income categories, three categories for the level of parents' education, dummy for physically active parents, number of siblings (categories: none, one, two, three or more), dummy for being a foreigner, birth weight, four categories for year of birth (1985 – 1990, 1991 – 1995, 1996 – 2000, 2001 – 2005), four categories of community size (<5,000, 5,000 – 20,000, 20,000 – 100,000, >100,000), dummy for East Germany, and educational spending per student.[19]

We estimate the effects for the five outcome groups separately. Observations with at least one missing value in the respective outcome group are excluded. Therefore, the number of observations in each outcome group can differ. The point estimates are robust to excluding all observation with at least one missing outcome in all groups, as we show in a sensitivity analysis. However, the estimates are less precise due to the smaller sample size. We thus favour the group-wise estimation. The cognitive skill outcome group excludes in addition students of the first and second grade because they usually do not receive grades. We observe a minority of

[18] Alternatives that also achieve perfect balancing are entropy balancing (Hainmueller, 2012), genetic matching (Diamond & Sekhon, 2013) and kernel balancing (Hazlett, 2016). However, IPT is locally efficient, double robust and has lower higher order bias than a large class of first-order equivalent alternative estimators (Graham, Pinto, & Egel, 2012). The automatic balancing property is an important advantage in our study because estimators that might be usually considered to be asymptotically more efficient (Huber, Lechner, & Wunsch, 2013), turned out to have difficulties in obtaining sufficient covariate balance in finite samples.

[19] Alternative coding of the categories affects the results only marginally. IPT provides a specification test for the propensity score. The results in the sensitivity section shows that this test does not reject the chosen model.

students reporting grades but they are most likely not representative for all students of the first two grades.

We ensure common support in each outcome group and subsample. Common support means that we have overlapping distributions of the propensity scores in both requirement groups. Overlap of the propensity score is achieved by running a first regression using all available observations and calculating the propensity scores for the groups with two and three lesson requirements, respectively. We restrict the final estimation sample to the observations where propensity scores overlap for the two and the three lesson requirement groups. This procedure results in the exclusion of at most 10% of the observations.[20]

Inference is based on 4999 weighted bootstraps (Barbe & Bertail, 1995). The bootstraps are clustered on the level of instrument variation, namely the state-school type-grade-wave level. Symmetric p-values are used to assess the statistical significance of the estimates.[21]

5 Descriptive statistics

This section provides descriptive statistics that assess the relation of required and actual PE lessons. A full description, mean values, and standard deviations of all variables used in the analysis are provided in Appendix G.

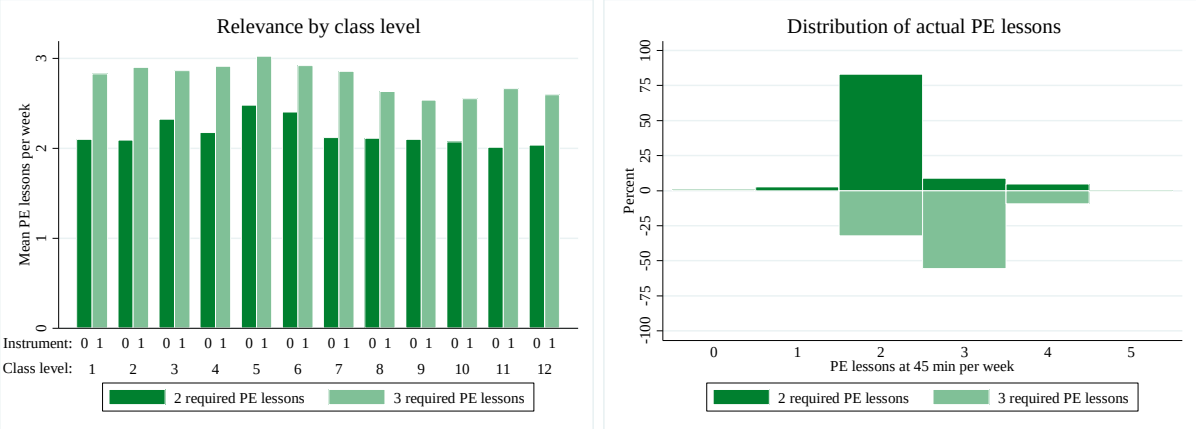
The identification strategy argues that the relevance of our instrument is not problematic in our application. The two graphs of Figure 5.1 provide evidence for this claim. Relevance means in our specific case that students with three required PE lessons report more PE lessons per week than students with two required lessons. The left graph of Figure 5.1 shows that this pattern is consistent and pronounced for all class levels. Students with three required lessons

[20] Again, a sensitivity check is provided showing that the results do not depend on the common support adjustment.

[21] P-values based on t-values (point estimate divided by the standard deviation of the bootstrap distribution) are nearly identical, which confirms our observation that the estimators are approximately normally distributed.

report on average 0.7 additional actual lessons compared to students with two required lessons (2.8 vs. 2.1 lessons). The magnitude of the differences vary across class levels but are highly statistically significant for each class level separately.

Figure 5.1: Required and actual PE lessons



The right graph of Figure 5.1 compares the distributions of actual PE lessons for the two and three lesson requirements. The graph explains the observed mean differences. Over 80% of the students with a requirement of two lessons report compliance with the curriculum. The compliance is considerably lower for the group with three required lessons. However, the majority of 56% of students still complies exactly with the curriculum. Another sizable fraction of 32% of the students receives only two lessons. This explains why the average difference is clearly below one, which would be expected if all students would receive the required number of PE lessons. However, compliance seems large enough to provide a strong and therefore relevant instrument. Further, the right graph shows that the estimated weighted LATE is mostly driven by students who are at schools that comply with the three lessons requirement.[22]

The descriptive analysis shows that the German institutional setting enables us to construct a powerful instrument. The feature that the minimum requirements are either two or

[22] If the three lessons requirement would only shift mass from two to three actual lessons, we would even estimate the standard LATE parameter. However, especially the fraction of four actual hours is significantly higher for those with three lesson requirements, which rules out this special case. Therefore, we cannot identify whether students are shifted, e.g., from two to three, two to four, or three to four hours and identify the weighted LATE as described above.

three provides a transparent and, combined with sufficient compliance, strong instrument. The previous studies for the US (Cawley et al., 2013, 2007; Dills et al., 2011; Sabia et al., 2016) need to deal with much more heterogeneous regulations, which complicates the construction of a strong instrument. The US studies need to make sense of states with no requirements for PE at all, states with required PE but without specified amount, and states with specific PE requirements. Therefore, it is not surprising that their resulting instruments are weak[23] and that their estimates are rather imprecise. The clear-cut German setting allows us to improve in this direction with F-statistics of the first stage exceeding 100.

6 Results

6.1 Main results

Our analysis comprises five outcome groups. The tables in Appendix E show the full set of outcome variables of each group for all observations, as well as for boys and girls separately. These tables also provide the outcome group specific *first stages*, which are highly significant with the lowest F-statistic being 179 for grades of girls. This confirms the descriptive evidence that required PE lessons are strong instruments for actual PE lessons. The number of actual PE lessons increases on average by 0.5 to 0.6 for students with three lesson requirements compared to those with two lesson requirements.

We account for multiple testing by calculating the joint significance of the effects in each outcome group-gender subsample. The associated F-statistics are shown in the last row of the tables in Appendix E. We interpret single effects only as significant if the F-statistic of the according outcome groups is significant at the 5%-level. This addresses the concern that some effects are significant by chance, if such a large number of outcomes is considered.

[23] Largest F-statistic being 33 in Cawley et al. (2007) for the subsample of girls.

Table 6.1 shows that we find significant and sizeable effects on *grades* of students. The average grade of math and German improves by 0.2 of a standard deviation (sd) considering all students. These findings suggest that more PE lessons can improve learning success in other subjects.

Table 6.1: Selected results

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Grades:						
German grade (std)	0.21***	0.06	0.13	0.09	0.25***	0.08
Math grade (std)	0.16**	0.06	0.16*	0.09	0.12	0.09
Average grade (std)	0.21***	0.06	0.17*	0.09	0.21**	0.09
Non-cognitive skills:						
Emotional symptoms abnormal (bin)	-0.03**	0.02	-0.01	0.02	-0.06**	0.02
Peer relations problems abnormal (bin)	0.01	0.01	0.05***	0.02	-0.03*	0.02
Total index abnormal (bin)	0.02*	0.01	0.04**	0.02	0.01	0.02
Motor skills:						
Side-steps (std)	0.09**	0.04	0.08	0.05	0.09	0.06
Balancing backwards (std)	0.12**	0.05	0.09	0.08	0.16**	0.07
Inserting pins duration (std)	-0.09**	0.04	-0.07	0.06	-0.09***	0.06
Stand and reach (std)	0.17***	0.05	0.07	0.07	0.28***	0.08
Physical activity:						
# of days with PE	0.84***	0.04	0.82***	0.04	0.85***	0.05
# of days active per week (w/o PE)	0.19*	0.10	0.14	0.14	0.21**	0.10
# of leisure sports	0.18***	0.06	0.12	0.08	0.24***	0.09
Health parameters:						
no individual and joint effects significant at 5%						

*Notes: This table summarises outcomes of the main results in Appendix E with at least one effect that is individually and jointly significant at the 5%-level. Standard error are based on 4999 weighted bootstraps clustered at state-school type-class-wave level. No. of observations vary for different outcome groups and are shown in Appendix E. * p<0.1, ** p<0.05, *** p<0.01*

The findings on *non-cognitive skills* are the most puzzling ones. Girls profit from PE by showing a significant reduction in *emotional symptoms*. However, more PE has adverse behavioural effects for boys. The probability that boys are classified as showing abnormal behaviour in the category *peer relations* problems increases by five percentage points. In

addition, the probability that the *total SDQ index* indicates abnormal behaviour increases significantly by 3.6 percentage points for boys.

The expected improvements in *motor skills* are mainly observed for girls. For both genders, we find improvements in the task *side-steps* (0.09 sd) that measures coordination and speed. Further improvements are concentrated among girls. They perform significantly better in *balancing exercises* (0.16 sd), are faster in the coordination task *inserting pins* (-0.09 sd), and have a higher *stretchability* of their body measured by the task *stand and reach* (0.28 sd).

Girls also drive the significant improvements in the outcome group *physical activity*. They report 0.2 more days for which they cumulate at least 60 minutes of moderate to vigorous physical activity outside of PE if they have one additional PE lesson. Further, they practise a larger variety of sports in their leisure time because the number of different leisure sports increases significantly. We find no evidence for the crowding out of extracurricular activities due to PE. The number of days per week with PE increases by 0.8 for all.

The effects for the considered *health parameters* are individually and jointly not significant. This is a surprising result and suggests that the training effect of 45 additional minutes of PE is not effective enough to be detectable in our data in the most obvious domain.

6.2 Effect heterogeneity

We investigate effect heterogeneity with regard to social status measured as *household income*. We compare the effects of PE across students from households with less and more than 2,500 € income per month.[24] As expected, the high income group performs better along all outcome groups. For example, the average grades between the two groups differ by 0.17 sd with average grades of -0.09 for low income and 0.08 for high income students.

[24] This threshold divides the sample in two subsamples of about the same size and coincides roughly with the median income in Germany.

Previous studies about school-based interventions show considerably larger effects for students with lower social status and suggest that they can help to decrease the achievement gap. We cannot confirm this for the effects of PE on non-cognitive and motor skills, active lifestyle, and health outcomes. The estimated effects for these outcomes differ only marginally across income groups and are mostly not statistically significant. However, we find substantial and significant differences in the effects on grades. The results in Appendix F show that high-income students experience no significant improvements in German and math grades. In contrast, low-income students improve their grades significantly up to 0.31 sd. The largest improvements are found for boys whose German and math grades increase by 0.30 and 0.36 sd, respectively. The advantage of splitting the sample by income is that the number of clusters that are available for estimation is reduced only moderately and thus only slightly affects the precision of the estimates. Other heterogeneity analyses are conducted by splitting the sample for East and West Germany or young and old. However, the precision of the estimates is substantially decreased due to a considerably smaller number of clusters. Therefore, we focus only on the results for income groups in the discussion below because this split was the only one that still led to sufficiently precise estimates.

6.3 Sensitivity analysis

We conduct a variety of robustness checks. Each robustness check reproduces the five tables of the main results while varying potentially critical features of the analysis. The corresponding tables are provided in Appendix H.

The first robustness check addresses the concern that *pre-school differences* between low and high PE states could drive our results. Table D.1 suggests that pre-school differences are mostly not significant for the available outcome variables. However, showing that the results are not sensitive to controlling for pre-school differences would strengthen the argument against policy endogeneity. We observe all outcomes except for grades, school-based PE, and push-

ups for pre-school children. Unfortunately, we observe these outcomes only in rare cases for the same individual. Therefore, we calculate state and wave specific means for the observed pre-school children and include these means as an additional control for the respective outcome. The estimated effects are nearly unaffected by controlling for pre-school differences. The effects vary within one standard error of the main results. The adverse effects on non-cognitive skills for boys are now even more pronounced and significant. The other statistically significant effects in the main results remain significant. The only exception is the positive effect on strength measured as side-steps that becomes insignificant after controlling for pre-school differences. We conclude that pre-school differences are not driving our main results.

Another concern could be that some *outliers* are responsible for the strong *first stage*. About 4% of the students report numbers of PE lessons smaller than two or larger than four. Further, low outliers are much more frequent for two-lesson requirements and high outliers are much more frequent for three-lesson requirements. We exclude all numbers of PE lessons below two and above four to check whether this correlation influences the result. The first stage is indeed about 0.1 lower if these outliers are excluded. However, the lowest F-statistic is still above 140. As expected, the effects are on average slightly higher due to the smaller denominator for the weighted LATE but all effects differ at most by one standard error from the main results.

The observations of the MoMo Study are not representative for the German population of interest because small states and different socio-economic groups are oversampled (Kamtsiuris, Lange, & Rosario, 2007). We ignored this fact so far because we included the relevant factors among our control variables. However, we rerun the analysis using the provided *sampling weights* that account for the sampling procedure and systematic non-response. Although standard the errors are about 50% larger than the ones in the main analysis, the

majority of the effects differ again only by at most one standard error from the main results. The few cases with larger deviations do not change the general conclusions.

The analysis of grades is conducted for students in *grade three and older*, while the other outcomes consider all students. We check whether restricting the estimation to only third grade and older affects the results in the other outcome categories. We find no notable differences to the main results besides the expected efficiency decreases due to fewer observations.

The main analysis considers different samples for the different outcome categories to avoid having missing outcome values decrease the sample size substantially. Restricting our sample to observations with valid entries in all considered outcomes creates a *balanced sample* with 3,420 observations for grades and 3,558 for the other four outcome groups. The effects are very similar to the main results, only the gender differences in the effects on non-cognitive skills are more pronounced.

The set of controls includes several variables that show no statistically significant differences for the two values of the instrument in Table C.1. Still, these variables are included in the set of controls for the main analysis because they are used in previous studies as well. We check the sensitivity of our results to the inclusion of the arguably irrelevant controls gender, physically active parents, siblings, and birth weight by estimating a *sparse model* containing only the statistically significant characteristics. Again, the main findings are robust to this change.

The *common support* adjustment does not affect the results and conclusions either. We run the analysis with the full sample without enforcing common support of the propensity score. The point estimates vary by less than half a standard error and the standard errors are only 5% larger or smaller. We conclude that common support considerations are of minor importance for our results.

Finally, we exploit a special feature of the IPT estimator to evaluate the specification of the propensity score. The IPT estimates two separate propensity scores for the two and the three lesson requirements groups. If the propensity score is correctly specified, the two estimated propensity scores should be identical. Therefore, we test the null-hypothesis that the coefficients in the two propensity score models are identical. We cannot reject the null-hypothesis with p-values of larger than 0.9 for all subsamples. This gives at least an indication that our propensity score is likely to be correctly specified.

7 Discussion

The previous section presents a variety of results on the five outcome categories of interest. This section discusses our findings with respect to the previous literature and potential explanations for the presented effects.

We find substantial positive effects on *cognitive skills* measured by German and math grades. This finding is in line with a variety of meta-studies that review the literature on the relationship of physical activity and academic achievements. The reviews of Trudeau and Shepard (2008), Singh et al. (2012), and Lees and Hopkins (2013) are most relevant for our study. They consider mostly quasi-experimental studies and conclude that increased PE has positive or neutral effects on academic achievements. This finding holds even for intervention studies where increased time in PE crowded-out instruction time in other subjects.[25] Such positive effects are probably not very surprising as such interventions are usually conducted by specially trained staff using modern methods of teaching and training. The two studies looking at regular PE in the US find mostly no significant effects and some positive effects on academic achievements (Cawley et al., 2013; Dills et al., 2011).

[25] A variety of other meta-studies documents positive or neutral effects of physical activity on academic achievements (Bird, Tripney, & Newman, 2013; Esteban-Cornejo et al., 2014; Howie & Pate, 2012; Rasberry et al., 2011).

Our results suggest that the same amount of German and math lessons is more productive for students with more PE lessons. A possible explanation is provided by a growing literature in neuroscience. Several meta-studies report that physical activity improves cognition, brain structure, and brain functions that are involved in attention, inhibition, and memory (Chaddock-Heyman, Hillman, Cohen, & Kramer, 2014; Hillman, Erickson, & Kramer, 2008; Verburgh, Königs, Scherder, & Oosterlaan, 2014). Most of these studies show this improvements directly after exercising (e.g. Hillman et al., 2009). This mirrors the average school day quite well where PE lessons are usually followed by lessons in other subjects. One additional PE lesson increases the days at school with PE on average by 0.8 days. This means that the improved brain functions materialise for more lessons per week, which could explain our positive findings.

We show that low-income children mainly drive the positive effects. This is in line with randomised control trials that investigate the effects of exercising on cognitive processes and find larger positive effects for low-income children (Tine, 2014; Tine & Butler, 2012).

The magnitude of the estimated gains of about 0.2 sd is similar to the effects of the participation in club sports estimated by Felfe, Lechner, and Steinmayr (2016) using the KiGGS dataset for Germany (0.13 – 0.25 sd). Comparisons of the magnitude with studies from other countries seem arbitrary because the grading system might not be comparable even after standardisation.

While the potentially positive effects of physical activity on cognitive skills is widely documented, the evidence regarding *non-cognitive skills* is rare and ambiguous. Only self-esteem is unambiguously increased by physical activity (Lees & Hopkins, 2013; Smith et al., 2014). Further, observational studies that also use the SDQ as outcomes tend to find significantly fewer behavioural problems for more active children (Ussher, Owen, Cook, & Whincup, 2007; Wiles et al., 2008). This might not hold necessarily for the specific case of PE. Sociologists and psychologists discuss the potential benefits of physical activity for non-

cognitive skills (Coakley, 2011; Gould & Carson, 2008; Holt & Neely, 2011). However, they emphasise that the non-cognitive benefits of physical activity depend largely on the circumstances and could have adverse effects.

The documented adverse effects on non-cognitive skills for boys are driven by the category peer relations problems but conduct problems are also increased. Children with abnormal conduct problems are those children who bully other children, while children with abnormal peer relations problems are those children who are bullied by other children. This suggests increased bullying within or around PE lessons. This results stand in stark contrast to the results of Felfe et al. (2016) who document favourable effects of club sports participation on conduct and peer problems measured on the same scale. The interesting question is what drives these differences. One potential explanation lies in the possibilities to self-select into different kinds of sports. The self-selection into a particular sports *club* is most likely driven by specific skills related to the particular sport or by friendship networks. In contrast, PE *school* lessons provide usually the same sports activity for every student. Some boys outperform other boys in the different sports, which could create tensions between the "losers" and the "winners". These tensions could be unloaded after PE at the schoolyard to adjust the pecking order again.

Our results suggest that *motor skills* are significantly improved through PE lessons for girls but not for boys. This is in line with the findings of Okely, Booth and Patterson (2001) who find that the positive relation of time spent in organised sports and good motor skills is larger for girls. However, it is difficult to determine the direction of causality in such studies and it is still an open question if better motor skills lead to more physical activity or vice versa, while a positive relationship is well-documented (Holfelder & Schott, 2014). Exploiting an exogenous difference in PE in our study allows us to claim that the improved motor skills are actually caused by this increase.

The training of motor skills during PE lessons should prepare and encourage students for *physical activity* in and out of school. The observation that girls with more PE lessons practice a larger variety of sport during their leisure time is thus in line with the positive effects of motor skills for girls. Further, girls report a higher number of days per week at which they are at least moderately physically active for 60 minutes or more excluding time spent in PE. Therefore, the improvements in motor skills of girls are likely achieved partly by PE lessons and partly by the more versatile and more frequent extracurricular activities.

In general, we find no evidence for any compensation of extracurricular activity induced by more time in PE. The ActivityStat Hypothesis brought forward by Rowland (1998) suggests that increased activity in one domain, in our case PE, should decrease activity in other domains, in our case voluntary PE and extracurricular activity. However, a recent meta-study shows that the ActivityStat Hypothesis is not convincingly supported in the empirical research so far (Gomersall, Rowlands, English, Maher, & Olds, 2013): Of the 22 identified studies that focus on potential compensation, 12 studies find support for compensation while 10 studies find no support for compensation. Again, these studies do not evaluate regular PE but rather some narrower interventions. Cawley et al. (2013) find also no convincing evidence that regular PE crowds out other activities for US students.[26]

More specifically for Germany, we find no crowding out of participation in club sports on the extensive and the intensive margin. Our results cannot rule out that such a compensation would occur if the mostly two and three lessons were increased to, say, daily lessons as we observe mostly students in two instead of three lessons. Thus, such extrapolations could be misleading.

[26] They check participation in nine different types of physical activities outside regular PE and find only two decreases being significant at the 10%-level.

Finally, we find no statistically significant effects on *health* outcomes. This is in line with a review on the effects of PE concluding that there is limited evidence for positive health effects of PE (Pate, O'Neill, & McIver, 2011), and Sabia et al. (2016) who find no effects on body weight.[27] In contrast to our results, Cawley et al. (2013) find a sizable reduction of the BMI, the prevalence of overweight, and obesity for boys. However, the prevalence of overweight is much larger in their sample compared to our sample (31% vs. 7% overweight).

8 Conclusion

This study examines the effect of regular PE on child development by exploiting variation across and within states in German PE requirements. It is the first study that comprehensively considers all five domains that are supposed to be positively affected by PE: cognitive skills, non-cognitive skills, motor skills, physical activity, and health. The majority of the effects show either statistically significant positive or insignificant effects on the targeted domains. Especially the significantly positive effects on grades suggest that PE can support the development of cognitive skills. Further, these positive effects are concentrated among low income children and indicate that more PE could be an effective measure to decrease educational inequality.

The findings of improvements in school grades make a strong case for the extension of PE. However, the substantially increased behavioural problems of boys show that there might be a cost to pay. The research design and data of this paper are not sufficient to detect the mechanisms responsible for the adverse effects on boys' non-cognitive skills. However, future research should aim to uncover the reasons for this finding to inform policymakers which characteristics of regular PE are responsible for this development.

[27] Tittlbach et al. (2010) find also no differences in health outcomes for students with more PE using the MoMo data of the Baseline and one-to-one matching based on age, gender, and social status.

Finally, in line with most of the previous studies, we are not able to detect any statistically significant improvements in health-related outcomes. Most likely, the variation of the intensity of physical activity in regular PE is not sufficient to create substantial effects in this domain. However, the improved motor skills of girls show that PE is effective in this, so far neglected, domain and that then improved skills arguably serve as a multiplier by encouraging them to engage in more frequent and more versatile extracurricular physical activities.

9 References

Barbe, P., & Bertail, P. (1995). *The Weighted Bootstrap* (Vol. 98). New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-1-4612-2532-4

Benjamin, R. M. (2010). The Surgeon General's Vision for a Healthy and Fit Nation. *Public Health Reports*, *125*(4), 514–515.

Bird, K. S., Tripney, J., & Newman, M. (2013). The educational impacts of young people's participation in organised sport: a systematic review. *Journal of Children's Services*, *8*(4), 264–275. https://doi.org/10.1108/JCS-04-2013-0014

Brettschneider, W.-D. (2005). *DSB Sprint-Studie: Eine Untersuchung zur Situation des Schulsports in Deutschland* (1st Edition). Aachen: Meyer & Meyer Sport.

Cawley, J., Frisvold, D., & Meyerhoefer, C. (2013). The impact of physical education on obesity among elementary school children. *Journal of Health Economics*, *32*(4), 743–755. https://doi.org/10.1016/j.jhealeco.2013.04.006

Cawley, J., Meyerhoefer, C., & Newhouse, D. (2007). The impact of state physical education requirements on youth physical activity and overweight. *Health Economics*, *16*(12), 1287–1301. https://doi.org/10.1002/hec.1218

Chaddock-Heyman, L., Hillman, C. H., Cohen, N. J., & Kramer, A. F. (2014). III. The importance of physical activity and aerobic fitness for cognitive control and memory in children. *Monographs of the Society for Research in Child Development*, *79*(4), 25–50. https://doi.org/10.1111/mono.12129

Coakley, J. (2011). Youth Sports: What Counts as "Positive Development?" *Journal of Sport & Social Issues*, 0193723511417311. https://doi.org/10.1177/0193723511417311

De Meester, F., van Lenthe, F. J., Spittaels, H., Lien, N., & De Bourdeaudhuij, I. (2009). Interventions for promoting physical activity among European teenagers: a systematic review. *The International Journal of Behavioral Nutrition and Physical Activity*, *6*, 82. https://doi.org/10.1186/1479-5868-6-82

Demetriou, Y., & Höner, O. (2012). Physical activity interventions in the school setting: A systematic review. *Psychology of Sport and Exercise*, *13*(2), 186–196. https://doi.org/10.1016/j.psychsport.2011.11.006

Demetriou, Y., Sudeck, G., Thiel, A., & Höner, O. (2015). The effects of school-based physical activity interventions on students' health-related fitness knowledge: A systematic review. *Educational Research Review*, *16*, 19–40. https://doi.org/10.1016/j.edurev.2015.07.002

Diamond, A., & Sekhon, J. S. (2013). Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics*, *95*(3), 932–945. https://doi.org/10.1162/REST_a_00318

Dills, A. K., Morgan, H. N., & Rotthoff, K. W. (2011). Recess, physical education, and elementary school student outcomes. *Economics of Education Review*, *30*(5), 889–900. https://doi.org/10.1016/j.econedurev.2011.04.011

Esteban-Cornejo, I., Gómez-Martínez, S., Tejero-González, C. M., Castillo, R., Lanza-Saiz, R., Vicente-Rodríguez, G., … Martinez-Gomez, D. (2014). Characteristics of extracurricular physical activity and cognitive performance in adolescents. The AVENA study. *Journal of Sports Sciences*, *32*(17), 1596–1603. https://doi.org/10.1080/02640414.2014.910607

Fairclough, S., & Stratton, G. (2005). Physical activity levels in middle and high school physical education: a review. *Pediatric Exercise Science*, 17(3), 217-236.

Felfe, C., Lechner, M., & Steinmayr, A. (2016). Sports and Child Development. PLOS ONE, 11(5), e0151729. https://doi.org/10.1371/journal.pone.0151729

Frölich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, *139*(1), 35–75.

Gomersall, S. R., Rowlands, A. V., English, C., Maher, C., & Olds, T. S. (2013). The ActivityStat hypothesis: the concept, the evidence and the methodologies. *Sports Medicine (Auckland, N.Z.)*, *43*(2), 135–149. https://doi.org/10.1007/s40279-012-0008-7

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A Research Note. *Journal of Child Psychology and Psychiatry*, *38*(5), 581–586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

Gould, D., & Carson, S. (2008). Life skills development through sport: current status and future directions. *International Review of Sport and Exercise Psychology*, *1*(1), 58–78. https://doi.org/10.1080/17509840701834573

Graham, B. S., Pinto, C. C. D. X., & Egel, D. (2012). Inverse Probability Tilting for Moment Condition Models with Missing Data. *The Review of Economic Studies*, *79*(3), 1053–1079. https://doi.org/10.1093/restud/rdr047

Hainmueller, J. (2012). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, *20*(1), 25–46. https://doi.org/10.1093/pan/mpr025

Harris, K. C., Kuramoto, L. K., Schulzer, M., & Retallack, J. E. (2009). Effect of school-based physical activity interventions on body mass index in children: a meta-analysis. *Canadian Medical Association Journal*, *180*(7), 719–726. https://doi.org/10.1503/cmaj.080966

Hazlett, C. (2016). Kernel Balancing: A flexible non-parametric weighting procedure for estimating causal effects. *arXiv:1605.00155 [Math, Stat]*. Retrieved from http://arxiv.org/abs/1605.00155

Hillman, C. H., Erickson, K. I., & Kramer, A. F. (2008). Be smart, exercise your heart: exercise effects on brain and cognition. *Nature Reviews Neuroscience*, *9*(1), 58–65. https://doi.org/10.1038/nrn2298

Hillman, C. H., Pontifex, M. B., Raine, L. B., Castelli, D. M., Hall, E. E., & Kramer, A. F. (2009). The effect of acute treadmill walking on cognitive control and academic achievement in preadolescent children. *Neuroscience*, *159*(3), 1044–1054. https://doi.org/10.1016/j.neuroscience.2009.01.057

Holfelder, B., & Schott, N. (2014). Relationship of fundamental movement skills and physical activity in children and adolescents: A systematic review. *Psychology of Sport and Exercise*, *15*(4), 382–391. https://doi.org/10.1016/j.psychsport.2014.03.005

Holt, N. L., & Neely, K. C. (2011). Positive Youth Development Through Sport: A Review. *Revista Iberoamericana de Psicología Del Ejercicio Y El Deporte*, *6*(2), 299–316.

Howie, E. K., & Pate, R. R. (2012). Physical activity and academic achievement in children: A historical perspective. *Journal of Sport and Health Science*, *1*(3), 160–169. https://doi.org/10.1016/j.jshs.2012.09.003

Huber, M., Lechner, M., & Wunsch, C. (2013). The performance of estimators based on the propensity score. *Journal of Econometrics*, *175*(1), 1–21. https://doi.org/10.1016/j.jeconom.2012.11.006

Hynynen, S.-T., Stralen, M. M. van, Sniehotta, F. F., Araújo-Soares, V., Hardeman, W., Chinapaw, M. J. M., … Hankonen, N. (2016). A systematic review of school-based interventions targeting physical activity and sedentary behaviour among older adolescents. *International Review of Sport and Exercise Psychology*, *9*(1), 22–44. https://doi.org/10.1080/1750984X.2015.1081706

Imbens, G. W., & Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, *62*(2), 467–475. https://doi.org/10.2307/2951620

Institute of Medicine of the National Academies. (2013). *Educating the Student Body: Taking Physical Activity and Physical Education to School* (Vol. H. W. Kohl III, H. D. Cook (Eds.)). Washington DC: The National Academies Press: National Academies Press.

Jekauc, D., Wagner, M. O., Kahlert, D., & Woll, A. (2013). Reliabilität und Validität des MoMo-Aktivitätsfragebogens für Jugendliche (MoMo-AFB). *Diagnostica*, *59*(2), 100–111. https://doi.org/10.1026/0012-1924/a000083

Kamtsiuris, P., Lange, M., & Rosario, A. S. (2007). Der Kinder- und Jugendgesundheitssurvey (KiGGS): Stichprobendesign, Response und Nonresponse-Analyse. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, *50*(5–6), 547–556. https://doi.org/10.1007/s00103-007-0215-9

Kriemler, S., Meyer, U., Martin, E., van Sluijs, E. M. F., Andersen, L. B., & Martin, B. W. (2011). Effect of school-based interventions on physical activity and fitness in children and adolescents: a review of reviews and systematic update. *British Journal of Sports Medicine*, *45*(11), 923–930. https://doi.org/10.1136/bjsports-2011-090186

Kurth, B.-M., Kamtsiuris, P., Hölling, H., Schlaud, M., Dölle, R., Ellert, U., … Wolf, U. (2008). The challenge of comprehensively mapping children's health in a nation-wide health survey: Design of the German KiGGS-Study. *BMC Public Health*, *8*(1), 196. https://doi.org/10.1186/1471-2458-8-196

Lavelle, H. V., Mackay, D. F., & Pell, J. P. (2012). Systematic review and meta-analysis of school-based interventions to reduce body mass index. *Journal of Public Health (Oxford, England)*, *34*(3), 360–369. https://doi.org/10.1093/pubmed/fdr116

Lees, C., & Hopkins, J. (2013). Effect of Aerobic Exercise on Cognition, Academic Achievement, and Psychosocial Function in Children: A Systematic Review of Randomized Control Trials. *Preventing Chronic Disease*, *10*. https://doi.org/10.5888/pcd10.130010

Lunn, P. D., & Kelly, E. (2015). Participation in School Sport and Post-School Pathways: Evidence from Ireland. *National Institute Economic Review*, *232*(1), R51–R66. https://doi.org/10.1177/002795011523200106

Møller, N. C., Tarp, J., Kamelarczyk, E. F., Brønd, J. C., Klakk, H., & Wedderkopp, N. (2014). Do extra compulsory physical education lessons mean more physically active children-findings from the

childhood health, activity, and motor performance school study Denmark (The CHAMPS-study DK). *International Journal of Behavioral Nutrition and Physical Activity*, 11(1), 121.

Okely, A. D., Booth, M. L., & Patterson, J. W. (2001). Relationship of physical activity to fundamental movement skills among adolescents. *Medicine and Science in Sports and Exercise*, *33*(11), 1899–1904.

Pate, R. R., O'Neill, J. R., & McIver, K. L. (2011). Physical Activity and Health: Does Physical Education Matter? *Quest*, *63*(1), 19–35. https://doi.org/10.1080/00336297.2011.10483660

Quitério, A. L. D. (2012). School physical education: The effectiveness of health-related interventions and recommendations for health-promotion practice. *Health Education Journal*, 0017896912460934. https://doi.org/10.1177/0017896912460934

Rasberry, C. N., Lee, S. M., Robin, L., Laris, B. A., Russell, L. A., Coyle, K. K., & Nihiser, A. J. (2011). The association between school-based physical activity, including physical education, and academic performance: A systematic review of the literature. *Preventive Medicine*, *52*, *Supplement*, S10–S20. https://doi.org/10.1016/j.ypmed.2011.01.027

Rowland, T. W. (1998). The biological basis of physical activity. *Medicine and Science in Sports and Exercise*, *30*(3), 392–399.

Sabia, J. J., Nguyen, T. T., & Rosenberg, O. (2016). High School Physical Education Requirements and Youth Body Weight: New Evidence from the YRBS. *Health Economics*. https://doi.org/10.1002/hec.3399

Schmidt, S., Will, N., Henn, A., Reimers, A., & Woll, A. (2016). Der Motorik-Modul Aktivitätsfragebogen (MoMo-AFB). Leitfaden zur Anwendung und Auswertung. Karlsruhe: KIT.

Singh, A., Uijtdewilligen, L., Twisk, J. W. R., van Mechelen, W., & Chinapaw, M. J. M. (2012). Physical activity and performance at school: A systematic review of the literature including a methodological quality assessment. *Archives of Pediatrics & Adolescent Medicine*, *166*(1), 49–55. https://doi.org/10.1001/archpediatrics.2011.716

Smith, J. J., Eather, N., Morgan, P. J., Plotnikoff, R. C., Faigenbaum, A. D., & Lubans, D. R. (2014). The health benefits of muscular fitness for children and adolescents: a systematic review and meta-analysis. *Sports Medicine (Auckland, N.Z.)*, *44*(9), 1209–1223. https://doi.org/10.1007/s40279-014-0196-4

Tine, M. T. (2014). Acute aerobic exercise: an intervention for the selective visual attention and reading comprehension of low-income adolescents. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00575

Tine, M. T., & Butler, A. G. (2012). Acute aerobic exercise impacts selective attention: an exceptional boost in lower-income children. *Educational Psychology*, *32*(7), 821–834. https://doi.org/10.1080/01443410.2012.723612

Tittlbach, S., Sygusch, R., Brehm, W., Seidel, I., & Bös, K. (2010). Sportunterricht : Gesundheitschance für inaktive Kinder und Jugendliche? *Sportwissenschaft (Heidelberg)*, *40*(2), S. 120-126.

Trudeau, F., & Shephard, R. J. (2008). Physical education, school physical activity, school sports and academic performance. *International Journal of Behavioral Nutrition and Physical Activity*, *5*(1), 10. https://doi.org/10.1186/1479-5868-5-10

UNESCO. (2014). *World-wide survey of school physical education*. Paris : United Nations educational, scientific and cultural organization. Retrieved from http://doc.rero.ch/record/255588

Ussher, M. H., Owen, C. G., Cook, D. G., & Whincup, P. H. (2007). The relationship between physical activity, sedentary behaviour and psychological wellbeing among adolescents. Social psychiatry and psychiatric epidemiology, 42(10), 851-856.

Verburgh, L., Königs, M., Scherder, E. J. A., & Oosterlaan, J. (2014). Physical exercise and executive functions in preadolescent children, adolescents and young adults: a meta-analysis. *British Journal of Sports Medicine*, *48*(12), 973–979. https://doi.org/10.1136/bjsports-2012-091441

Wagner, M. O., Bös, K., Jekauc, D., Karger, C., Mewes, N., Oberger, J., … Woll, A. (2014). Cohort profile: the Motorik-Modul Longitudinal Study: physical fitness and physical activity as determinants of health development in German children and adolescents. *International Journal of Epidemiology*, *43*(5), 1410–1416. https://doi.org/10.1093/ije/dyt098

Wiles, N. J., Jones, G. T., Haase, A. M., Lawlor, D. A., Macfarlane, G. J., & Lewis, G. (2008). Physical activity and emotional problems amongst adolescents. Social psychiatry and psychiatric epidemiology, 43(10), 765.

WHO. (2010). *Global recommendations on physical activity for health*. Retrieved from http://www.who.int/dietphysicalactivity/publications/9789241599979/en/

Woll, A., Kurth, B.-M., Opper, E., Worth, A., & Bös, K. (2011). The “Motorik-Modul” (MoMo): physical fitness and physical activity in German children and adolescents. *European Journal of Pediatrics*, *170*(9), 1129–1142. https://doi.org/10.1007/s00431-010-1391-4

Appendices

A: Details on required PE lessons

A.1: Coding of required PE lessons

This Appendix shows how the required PE lessons are coded for the Baseline and Wave 1. Unfortunately, some curricula allow no direct assignment of required PE lessons to students based on their state, school type, and class. Three general issues arise and we deal with them in the following way:

- Instead of a single number, the curriculum states a range of required PE lessons. In these cases we assign the minimum required PE lessons. This is in line with Brettschneider (2005) who observes that most schools provide only the minimum amount of PE. We observe the same in our data.
- Some states rely on so-called *Kontingentstundentafeln* (contingency curricula) that specify a total number of required hours for several class levels. E.g., 12 lessons in grades 1 to 4. In such a case the lessons can be uniformly distributed, which is what we do by assigning three hours to each class level. If a uniform split is not possible with, e.g., 17 lessons for grades 5 to 10, we assign three hours to the lower grades (5 to 9) and two hours to the highest grade of the range. This procedure is in line with the empirical observation for these cases.
- Some curricula state a specific number of lessons only for a combination of subjects like PE, arts, and music combined. We are not able to assign a specific value in these cases. Thus, the corresponding students are not considered in the analysis.

Table A.1.1: PE requirements by states, school type, and grade - Baseline

Baseline (2004-2006)														
		Class level												
State	School type	1	2	3	4	5	6	7	8	9	10	11	12	13
Baden-Württemberg	Primary school	3	3	3	3	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	3	3	3	3	3	2	-	-	-
	Academic track	-	-	-	-	3	3	3	3	2	2	2	2	2
	Comprehensive school	-	-	-	-	-	-	-	-	-	-	-	-	-
Bavaria	Primary school	2	3	3	3	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	2	2	2	2	2	2	-	-	-
	Academic track	-	-	-	-	2	2	2	2	2	2	2	2	2
	Comprehensive school	-	-	-	-	-	-	-	-	-	-	-	-	-
Berlin	Primary school	3	3	3	3	3	3	-	-	-	-	-	-	-

Baseline (2004-2006)														
		Class level												
	Basic / Intermediate track	-	-	-	-	-	-	3	3	3	3	-	-	-
	Academic track	-	-	-	-	-	-	3	3	3	3	2	2	-
	Comprehensive school	-	-	-	-	-	-	3	3	3	3	2	2	-
Brandenburg	Primary school	3	3	3	3	3	3	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	-	-	3	3	3	3	-	-	-
	Academic track	-	-	-	-	-	-	3	3	3	3	3	3	3
	Comprehensive school	-	-	-	-	-	-	3	3	3	3	3	3	3
Bremen	Primary school	Contingency curricula aesthetic education combining arts, music, and PE.												
	Basic / Intermediate track													
	Academic track													
	Comprehensive school													
Hamburg	Primary school	3	3	3	3	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	3	3	3	3	3	3	-	-	-
	Academic track	-	-	-	-	3	3	3	3	3	3	2	2	2
	Comprehensive school	-	-	-	-	3	3	3	3	3	3	2	2	2
Hesse	Primary school	3	3	3	3	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	3	3	3	3	2	2	-	-	-
	Academic track	-	-	-	-	3	3	3	3	2	2	2	2	2
	Comprehensive school	-	-	-	-	3	3	3	3	2	2	2	2	2
Lower Saxony	Primary school	2	2	2	2	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	2	2	2	2	2	2	-	-	-
	Academic track	-	-	-	-	2	2	2	2	2	2	2	2	2
	Comprehensive school	-	-	-	-	2	2	2	2	2	2	2	2	2
Mecklenburg-West Pomerania	Primary school	2	3	3	3	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	3	3	2	2	2	2	-	-	-
	Academic track	-	-	-	-	3	3	2	2	2	2	2	2	2
	Comprehensive school	-	-	-	-	3	3	2	2	2	2	2	2	2
North Rhine-Westphalia	Primary school	3	3	3	3	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	3	3	3	3	3	3	-	-	-
	Academic track	-	-	-	-	3	3	3	3	3	3	3	3	3
	Comprehensive school	-	-	-	-	3	3	3	3	3	3	3	3	3
Rhineland-Palatinate	Primary school	Contingency curricula combining arts, music, and PE.				-	-	-	-	-	-	-	-	-
	Basic / Intermediate track					3	3	3	3	2	2	-	-	-
	Academic track					3	3	3	3	2	2	2	2	2
	Comprehensive school					3	3	2	2	2	2	2	2	2
Saarland	Primary school	2	2	2	2	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	2	2	2	2	2	2	-	-	-
	Academic track	-	-	-	-	2	2	2	2	2	2	2	2	2
	Comprehensive school	-	-	-	-	2	2	2	2	2	2	2	2	2
Saxony	Primary school	3	3	3	3	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	3	3	3	3	3	3	-	-	-
	Academic track	-	-	-	-	3	3	3	2	2	2	2	2	2
	Comprehensive school	-	-	-	-	-	-	-	-	-	-	-	-	-
Saxony-Anhalt	Primary school	2	2	2	2	-	-	-	-	-	-	-	-	-

Baseline (2004-2006)														
		Class level												
	Basic / Intermediate track	-	-	-	-	3	3	2	2	2	2	-	-	-
	Academic track	-	-	-	-	3	3	2	2	2	2	2	2	2
	Comprehensive school	-	-	-	-	-	-	-	-	-	-	-	-	-
Schleswig-Holstein	Primary school	2	2	3	3	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	3	3	3	3	3	3	-	-	-
	Academic track	-	-	-	-	3	3	3	3	3	3	2	2	3
	Comprehensive school	-	-	-	-	-	-	-	-	-	-	-	-	-
Thuringia	Primary school	2	2	3	3	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	3	3	3	3	3	3	-	-	-
	Academic track	-	-	-	-	3	3	3	3	3	3	2	2	2
	Comprehensive school	-	-	-	-	3	3	3	3	3	3	2	2	2

Table A.1.2: PE requirements by states, school type, and grade – Wave 1

Wave 1 (2009-2012)														
		Class level												
States	School type	1	2	3	4	5	6	7	8	9	10	11	12	13
Baden-Württemberg	Primary school	3	3	3	3	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	3	3	3	3	3	2	-	-	-
	Academic track	-	-	-	-	3	3	3	3	2	2	2	2	2
	Comprehensive school	-	-	-	-	-	-	-	-	-	-	-	-	-
Bavaria	Primary school	2	3	3	3	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	2	2	2	2	2	2	-	-	-
	Academic track[1)]	-	-	-	-	3	3	3	2	2	2	2	2	2
	Comprehensive school	-	-	-	-	-	-	-	-	-	-	-	-	-
Berlin	Primary school	3	3	3	3	3	3	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	-	-	3	3	3	3	-	-	-
	Academic track	-	-	-	-	-	-	3	3	3	3	2	2	-
	Comprehensive school	-	-	-	-	-	-	3	3	3	3	2	2	-
Brandenburg	Primary school	3	3	3	3	3	3	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	-	-	3	3	3	3	-	-	-
	Academic track	-	-	-	-	-	-	3	3	3	3	3	3	3
	Comprehensive school	-	-	-	-	-	-	3	3	3	3	3	3	3
Bremen	Primary school	Contingency curricula aesthetic education combining arts, music, and PE.												
	Basic / Intermediate track													
	Academic track													
	Comprehensive school													
Hamburg	Primary school	3	3	3	3	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	3	3	3	3	3	3	-	-	-
	Academic track	-	-	-	-	3	3	3	3	3	3	2	2	2
	Comprehensive school	-	-	-	-	3	3	3	3	3	3	2	2	2
Hesse	Primary school	3	3	3	3	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	3	3	3	3	2	2	-	-	-
	Academic track	-	-	-	-	3	3	3	3	2	2	2	2	2
	Comprehensive school	-	-	-	-	3	3	3	3	2	2	2	2	2
Lower Saxony	Primary school	2	2	2	2	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	2	2	2	2	2	2	-	-	-
	Academic track	-	-	-	-	2	2	2	2	2	2	2	2	2

Wave 1 (2009-2012)														
		Class level												
	Comprehensive school	-	-	-	-	2	2	2	2	2	2	2	2	2
Mecklenburg-West Pomerania	Primary school	2	3	3	3	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	3	3	2	2	2	2	-	-	-
	Academic track	-	-	-	-	3	3	2	2	2	2	2	2	2
	Comprehensive school	-	-	-	-	3	3	2	2	2	2	2	2	2
North Rhine-Westphalia	Primary school	3	3	3	3	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	3	3	3	3	3	3	-	-	-
	Academic track	-	-	-	-	3	3	3	3	3	3	3	3	3
	Comprehensive school	-	-	-	-	3	3	3	3	3	3	3	3	3
Rhineland-Palatinate	Primary school	Contingency curricula combining arts, music, and PE.				-	-	-	-	-	-	-	-	-
	Basic / Intermediate track					3	3	3	3	2	2	-	-	-
	Academic track					3	3	3	3	2	2	2	2	2
	Comprehensive school					3	3	2	2	2	2	2	2	2
Saarland	Primary school	2	2	2	2	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	2	2	2	2	2	2	-	-	-
	Academic track	-	-	-	-	2	2	2	2	2	2	2	2	2
	Comprehensive school	-	-	-	-	2	2	2	2	2	2	2	2	2
Saxony	Primary school	3	3	3	3	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	3	3	3	3	3	3	-	-	-
	Academic track	-	-	-	-	3	3	3	2	2	2	2	2	2
	Comprehensive school	-	-	-	-	-	-	-	-	-	-	-	-	-
Saxony-Anhalt	Primary school	2	2	2	2	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	3	3	2	2	2	2	-	-	-
	Academic track	-	-	-	-	3	3	2	2	2	2	2	2	2
	Comprehensive school	-	-	-	-	-	-	-	-	-	-	-	-	-
Schleswig-Holstein	Primary school	Contingency curricula aesthetic education combining arts, music, and PE (changed in 2007).												
	Basic / Intermediate track													
	Academic track													
	Comprehensive school													
Thuringia	Primary school	2	2	3	3	-	-	-	-	-	-	-	-	-
	Basic / Intermediate track	-	-	-	-	3	3	3	3	3	3	-	-	-
	Academic track	-	-	-	-	3	3	3	3	3	3	2	2	2
	Comprehensive school	-	-	-	-	3	3	3	3	3	3	2	2	2

Notes: Shaded in grey are changes to the Baseline.

1) Changed in 2007

A.2: Relation of PE lessons and other subjects

This section complements the graphical illustrations in Section 2. It investigates whether more PE lessons crowd out other subjects or increase total time at school. To this end, we calculate the average required lessons in all subjects over the first ten years. Then we regress the average of each subject on the average PE lessons and school type dummies. Table A.2.1 confirms the graphical finding that total time is very stable for different PE requirements but especially electives are crowded out. However, given that we only observe 33 different school tracks, the statistical power of this analysis is very limited.

Table A.2.1: Relation of PE lessons and other subjects

Relation of average PE lessons and	Average PE		Average PE > 2.5	
average instruction time in ...	Coef.	S.E.	Coef.	S.E.
... PE	-	-	0.66***	0.06
... German	0.17	0.30	0.12	0.22
... math	-0.29	0.23	-0.23	0.16
... foreign languages	0.25	0.24	0.18	0.17
... religion	-0.33	0.22	0.04	0.17
... music	0.13	0.08	0.12*	0.06
... arts	-0.03	0.08	0.02	0.06
... nature	-0.25	0.19	-0.07	0.14
... social	-0.25	0.26	-0.22	0.19
... electives	-0.31	0.74	-0.77	0.52
... total	0.10	0.51	-0.15	0.37
School type dummies	X		X	
# of observations	33		33	

*Notes: Each coefficient comes from a separate OLS regression that regresses a measure for average time required in other subjects on the average required PE lessons. The first column uses the simple average entering linearly and the second column a dummy for average PE larger than 2.5. Dummies for school types are always included. * p<0.1, ** p<0.05, *** p<0.01*

A.3 Average PE requirements across states

Figure A.3.1 shows the average values of PE requirements by state, calculated as the mean in the subsample of students living in the specific state (Table A.3.1 shows the respective numbers).

The dark blue states are Lower Saxony and Saarland with required PE lessons of two for all grades. The light blue states are North-Rhine-Westphalia and Brandenburg with required PE lessons of three for all grades. The three small grey parts are the city states Berlin, Bremen and Hamburg that are excluded from the analysis. The averages for the rest of the states are somewhere between 2 and 3, as indicated by the particular intensity of the colour.

Figure A.3.1: Average PE requirements across states

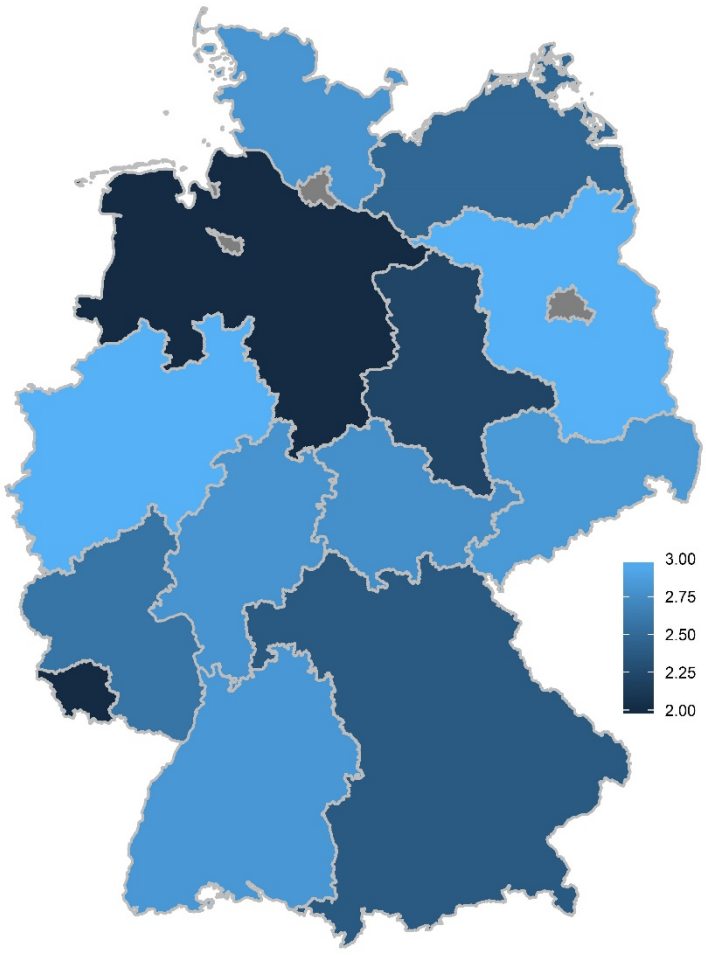


Table A.3.1: Average PE requirements across states

State	Mean PE requirement	# of obs.
Baden-Württemberg	2.83	703
Bavaria	2.37	946
Brandenburg	3.00	389
Hesse	2.80	235
Lower Saxony	2.00	470
Mecklenburg-West Pomerania	2.47	266
North Rhine-Westphalia	3.00	1,031
Rhineland-Palatinate	2.58	167
Saarland	2.00	41
Saxony	2.84	512
Saxony-Anhalt	2.21	275
Schleswig-Holstein	2.81	77
Thuringina	2.77	311
Total	2.66	5,423

Notes: This table shows the means and number of observations underlying Figure A.3.1.

B: Data

Table B.1 explains in detail the sampling procedure and the construction of the final sample. More details about the sampling procedure are provided in Kamtsiuris, Lange, and Rosario (2007).

Table B.1: Sampling procedure and data preparation

Sampling procedure:

Baseline (2003 – 2006):
167 sampling points in Germany for KiGGS study
28,400 invited to participate
17,641 participate in KiGGS
7,866 randomly assigned to MoMo
4,529 participate in MoMo Baseline
Wave 1 (2009 – 2012):
2,842 of 4,529 are observed again
2,317 newly recruited
Cross-section of 5,159 in wave 1

Preparation of data:	# of Obs.
Pooled sample	9688
- Children out of school	-3314
- Students without well-defined requirements:	
- Bremen	-32
- Schleswig-Holstein wave 1	-103
- Rhineland-Palatine primary school	-106
Raw sample	6133
- Missing or not plausible # of PE lessons (>10)	-305
- City states	-175
- Missing controls	-230
Final sample size	5423
Unique individuals	4698

C: Selection into higher PE requirements

The following table shows selection into the three lessons requirement.

Table C.1: Selection into higher PE requirements

	Mean comparison		Logit	
	3 lessons	2 lessons	AME	S.E.
Class level 1	0.06	0.11	*Reference categorie*	
Class level 2	0.11	0.06	0.23	0.13
Class level 3	0.13	0.04	0.40**	0.14
Class level 4	0.12	0.04	0.38**	0.14
Class level 5	0.12	0.06	0.36**	0.13
Class level 6	0.12	0.07	0.37**	0.13
Class level 7	0.11	0.10	0.27*	0.12
Class level 8	0.08	0.12	0.20	0.13
Class level 9	0.07	0.15	0.12	0.13
Class level 10	0.05	0.13	0.12	0.13
Class level 11	0.02	0.09	0.03	0.17
Class level 12	0.01	0.03	0.07	0.16
Primary school	0.43	0.25	*Reference categorie*	
Basic / Intermediate school	0.25	0.29	*Reference categorie*	
Academic track	0.26	0.42	-0.05	0.06
Comprehensive school	0.06	0.04	0.04	0.08
HH income < 1,000€	0.11	0.11	*Reference categorie*	
HH income 1,000 - 1,500€	0.08	0.07	-0.01	0.03
HH income 1,500 - 2,000€	0.12	0.11	-0.03	0.03
HH income 2,000 - 2,500€	0.18	0.18	-0.02	0.02
HH income 2,500 - 3,000€	0.18	0.17	-0.002	0.02
HH income 3,000 - 4,000€	0.19	0.18	0.02	0.02
HH income 4,000 - 5,000€	0.09	0.11	-0.01	0.03
HH income > 5,000€	0.05	0.06	0.01	0.03
Low education HH	0.10	0.13	*Reference categorie*	
Middle education HH	0.59	0.54	0.06*	0.02
High education HH	0.31	0.33	0.05*	0.03
Parents physically active	0.25	0.27	-0.01	0.01
Foreigner	0.03	0.02	0.12***	0.03
No siblings	0.14	0.14	*Reference categorie*	
One sibling	0.49	0.50	0.03	0.02
Two siblings	0.25	0.24	0.04	0.02
Three or more siblings	0.12	0.12	0.03	0.02
Birthweight in kilogram	3.39	3.37	-0.001	0.01
Cohort of 1985 - 1990	0.09	0.20	*Reference categorie*	
Cohort of 1991 - 1995	0.32	0.35	0.05	0.06
Cohort of 1996 - 2000	0.40	0.35	0.13*	0.07
Cohort of 2001 - 2005	0.18	0.10	0.27*	0.11
Female	0.50	0.51	-0.01	0.01

Table continues on next page >

	Mean comparison		Logit	
	3 lessons	2 lessons	AME	S.E.
East Germany	0.35	0.27	0.19***	0.05
< 5,000 inhabitans	0.23	0.28	*Reference categorie*	
5,000 - 20,000 inhabitans	0.32	0.33	0.04*	0.02
20,000 - 100,000 inhabitans	0.32	0.25	0.09**	0.03
>100,000 inhabitans	0.13	0.13	0.05	0.03
Educ. exp. per student in 100€	56.57	58.10	-0.02***	0.003
Observations	3,528	1,895	5,423	
# of clusters	-	-	498	

Notes: Average marginal effects of a logit regression are reported. The outcome variable is a binary indicator for three required PE lessons. Clustered standard errors in parentheses (p<0.1, ** p<0.05, *** p<0.01).*

D: Pre-school differences

Complementing the discussion about policy endogeneity in Section 4.1, Table D.1 shows only few significant differences in outcomes measured for 4 and 5 year old children that are not at school. Consequently, they should not yet be affected by the required experience, unless policy endogeneity is a concern.

Table D.1: Pre-school differences of high and low PE states

	Mean difference		IPT	
	Diff.	S.E.	Diff.	S.E.
Non-cognitive skils:				
Emotional symptoms index (std)	-0.03	0.07	-0.02	0.07
Conduct problems index (std)	-0.04	0.07	-0.03	0.07
Hyperactivity index (std)	0.08	0.07	0.07	0.07
Peer relations problems index (std)	0.13*	0.07	0.10	0.07
Asocial behaviour index (std)	0.07	0.07	0.05	0.07
Total index (std)	0.06	0.07	0.05	0.07
Emotional symptoms borderline or abnormal (bin)	0.02	0.02	0.03	0.02
Emotional symptoms abnormal (bin)	-0.01	0.02	0.003	0.02
Conduct problems borderline or abnormal (bin)	-0.01	0.04	-0.01	0.03
Conduct problems abnormal (bin)	-0.01	0.03	-0.001	0.03
Hyperactivity borderline or abnormal (bin)	0.03	0.03	0.03	0.03
Hyperactivity abnormal (bin)	-0.01	0.02	-0.01	0.02
Peer relations problems borderline or abnormal (bin)	0.04	0.03	0.03	0.03
Peer relations problems abnormal (bin)	0.03	0.02	0.02	0.02
Asocial behaviour borderline or abnormal (bin)	0.04**	0.02	0.05**	0.02
Asocial behaviour abnormal (bin)	0.03***	0.01	0.03***	0.01
Total index borderline or abnormal (bin)	0.02	0.02	0.03	0.02
Total index abnormal (bin)	-0.02	0.02	-0.02	0.02
Extracurricular activity:				
Physical activity in club sports in minutes	-3.45	3.08	-2.94	2.85
Physical activity in leisure sports in minutes	-5.25	6.65	-5.27	6.14
Physical activity in out of school in minutes	-8.70	7.33	-8.21	6.76
Participation club sports (bin)	-0.06	0.03	-0.05	0.03
# of days active per week	0.08	0.14	0.19	0.14
Compliance with WHO guideline (bin)	0.04	0.03	0.06*	0.03
Media consumption hrs/week	-0.47	0.46	-0.66	0.42
# of club sports	-0.05	0.05	-0.05	0.05
# of leisure sports	-0.05	0.07	-0.05	0.07
Health parameters:				
BMI	0.15	0.11	0.13	0.11
BMI (std)	0.09	0.07	0.08	0.07
Overweight (bin)	0.002	0.002	0.002	0.002

Table continues on next page >

	Mean difference		IPT	
	Diff.	S.E.	Diff.	S.E.
Health parameters (continued):				
Weight in kg	0.22	0.24	0.07	0.22
Weight in kg (std)	0.07	0.07	0.02	0.07
Subjective health 1-5	-0.01	0.04	-0.01	0.04
Subjective health good (bin)	0.004	0.03	0.02	0.03
Subjective health very good (bin)	-0.003	0.03	-0.01	0.03
Resting heart rate	0.06	0.85	0.09	0.80
Resting heart rate (std)	0.01	0.07	0.01	0.07
Height in cm	0.10	0.45	-0.24	0.40
Height (std)	0.02	0.07	-0.04	0.06
# of Observations	799			
Motor skills:				
Side-steps (std)	0.06	0.08	-0.001	0.078
Static stand (std)	0.10	0.08	0.064	0.082
Standing long jump (std)	-0.11	0.08	-0.124	0.080
Reaction test (std)	0.12	0.08	0.137*	0.078
Balancing backwards (std)	0.003	0.08	-0.010	0.079
Tracing lines mistakes (std)	0.13	0.08	0.091	0.081
Line tracking mistake duration (std)	0.08	0.08	0.069	0.083
Line tracking duration (std)	0.11	0.08	0.058	0.081
Inserting pins duration (std)	0.03	0.08	0.043	0.080
Stand and reach (std)	0.04	0.08	0.052	0.083
# of Observations	628			

*Notes: Mean difference between high and low PE states for 4 and 5 year old children. First, unconditional mean comparison. Second, Inverse Probability Tilting to control for household income, household composition, parents education, parents physical activity, foreign status, year of birth, East Germany, urbanisation, and education expenditure per student in the states. Push-ups are not available for this age group. Grades and school-based physical activities also not observed for pre-school children Standard errors obtained from 4999 weighted bootstraps. Inference based on symmetric p-values. * p<0.1, ** p<0.05, *** p<0.01*

E: Full main results

Table E.1: Main results - grades

	All		Boys		Girls		Boys - Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Diff.	S.E.
1st stage	0.53***	0.03	0.56***	0.04	0.50***	0.04	0.05	0.05
German grade (std)	0.21***	0.06	0.13	0.09	0.25***	0.08	-0.12	0.12
Math grade (std)	0.16**	0.06	0.16*	0.09	0.12	0.09	0.04	0.13
Average grade (std)	0.21***	0.06	0.17*	0.09	0.21**	0.09	-0.04	0.13
# of observations	4035		2055		1967			
# of clusters	443		392		388			
# of observations off support	284		83		214			
F-statistic of first stage	236.7***		201.6***		178.6***			
F-statistic for joint significane of LATE's	4.4***		1.1		3.0**			

*Notes: This table shows the weighted LATE estimates for students of the 3rd to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table E.2: Main results – non-cognitive skills

	All		Boys		Girls		Boys - Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Diff.	S.E.
1st stage	0.58***	0.03	0.60***	0.03	0.56***	0.03	0.05	0.05
Emotional symptoms index (std)	-0.04	0.05	-0.01	0.06	-0.07	0.08	0.06	0.10
Conduct problems index (std)	0.12**	0.05	0.15**	0.06	0.09	0.07	0.06	0.09
Hyperactivity index (std)	-0.05	0.04	-0.10	0.07	0.01	0.06	-0.11	0.09
Peer relations problems index (std)	0.09*	0.05	0.24***	0.07	-0.05	0.07	0.28***	0.10
Asocial behaviour index (std)	-0.06	0.05	-0.11	0.07	-0.01	0.07	-0.10	0.10
Total index (std)	0.03	0.05	0.07	0.06	-0.01	0.07	0.08	0.09
Emotional symptoms borderline or abnormal (bin)	-0.02	0.02	0.01	0.02	-0.04	0.03	0.05	0.04
Emotional symptoms abnormal (bin)	-0.03**	0.02	-0.01	0.02	-0.06**	0.02	0.05*	0.03
Conduct problems borderline or abnormal (bin)	0.04*	0.02	0.05*	0.03	0.03	0.03	0.02	0.04
Conduct problems abnormal (bin)	0.02	0.02	0.04*	0.02	0.01	0.02	0.03	0.03
Hyperactivity borderline or abnormal (bin)	0.01	0.02	0.01	0.03	0.02	0.02	-0.01	0.03
Hyperactivity abnormal (bin)	0.003	0.01	0.01	0.02	-0.01	0.02	0.02	0.03
Peer relations problems borderline or abnormal (bin)	0.02	0.02	0.08***	0.03	-0.04*	0.02	0.12***	0.03
Peer relations problems abnormal (bin)	0.01	0.01	0.05***	0.02	-0.03*	0.02	0.08***	0.03
Asocial behaviour borderline or abnormal (bin)	-0.01	0.01	0.01	0.02	-0.02	0.02	0.03	0.03
Asocial behaviour abnormal (bin)	-0.01	0.01	-0.01	0.01	-0.02*	0.01	0.01	0.02
Total index borderline or abnormal (bin)	0.02	0.02	0.06***	0.02	-0.01	0.02	0.07**	0.03
Total index abnormal (bin)	0.02*	0.01	0.04**	0.02	0.01	0.02	0.03	0.02
# of observations	5055		2580		2458			
# of clusters	494		440		434			
# of observations off support	364		121		260			
F-statistic of first stage	384.9***		313.6***		274.2***			
F-statistic for joint significane of LATE's	2.1***		2.7***		1.6**			

*Notes: This table shows weighted LATE estimates. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table E.3: Main results – motor skills

	All		Boys		Girls		Boys - Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Diff.	S.E.
1st stage	0.57***	0.03	0.58***	0.04	0.55***	0.04	0.04	0.05
Push-ups (std)	-0.09	0.05	-0.08	0.08	-0.12*	0.07	0.05	0.11
Side-steps (std)	0.09**	0.04	0.08	0.05	0.09	0.06	-0.01	0.08
Static stand (std)	0.02	0.05	0.02	0.08	0.03	0.07	-0.002	0.11
Standing long jum (std)	0.02	0.04	0.04	0.05	0.002	0.06	0.04	0.08
Reaction time (std)	0.05	0.05	0.04	0.06	0.05	0.07	-0.02	0.09
Balancing backwards (std)	0.12**	0.05	0.09	0.08	0.16**	0.07	-0.08	0.10
Line tracking mistakes (std)	-0.03	0.05	-0.05	0.08	0.01	0.07	-0.06	0.10
Line tracking mistake duration (std)	-0.04	0.06	-0.03	0.09	-0.05	0.06	0.03	0.11
Line tracking duration (std)	0.12*	0.07	0.08	0.09	0.14	0.09	-0.07	0.13
Inserting pins duration (std)	-0.09**	0.04	-0.07	0.06	-0.09	0.06	0.02	0.08
Stand and reach (std)	0.17***	0.05	0.07	0.07	0.28***	0.08	-0.21*	0.11
# of observations	4312		2226		2071			
# of clusters	482		430		420			
# of observations off support	293		92		216			
F-statistic of first stage	343.4***		240.1***		228.8***			
F-statistic for joint significane of LATE's	3.1***		1.1		2.5***			

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table E.4: Main results – physical activity

	All		Boys		Girls		Boys - Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Diff.	S.E.
1st stage	0.59***	0.03	0.60***	0.04	0.57***	0.04	0.03	0.05
School based:								
# of days with PE	0.84***	0.04	0.82***	0.04	0.85***	0.05	-0.03	0.06
# of voluntary PE lessons	-0.02	0.06	-0.02	0.09	-0.02	0.07	0.003	0.12
F-statistic for joint significane of LATE's	280.4***	0.01	206.7***	0.02	179.6***	0.01		
Extracurricular:								
Physical activity in club sports in minutes	-1.82	3.20	0.76	4.32	-3.32	4.73	4.07	6.40
Physical activity in leisure sports in minutes	4.06	6.39	8.54	9.46	-0.12	8.09	8.65	12.45
Physical activity out of school in minutes	2.24	7.28	9.29	10.62	-3.43	9.25	12.72	14.08
Participation club sports (bin)	-0.023	0.02	-0.03	0.03	-0.02	0.04	-0.01	0.05
# of days active per week	0.19*	0.10	0.14	0.14	0.21**	0.10	-0.07	0.18
Compliance with WHO guideline (bin)	-0.01	0.02	-0.02	0.03	0.001	0.02	-0.02	0.03
Media consumption hrs/week	0.67	0.68	1.05	0.99	0.52	0.83	0.54	1.29
# of club sports	-0.03	0.04	0.02	0.05	-0.06	0.06	0.08	0.08
# of leisure sports	0.18***	0.06	0.12	0.08	0.24***	0.09	-0.12	0.12
# of observations	4729		2394		2391			
# of clusters	484		429		429			
# of observations off support	339		123		160			
F-statistic of first stage	378.2***		287.1***		271.7***			
F-statistic for joint significane of LATE's	3.0***		1.6		2.2**			

Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level.
** p<0.1, ** p<0.05, *** p<0.01*

Table E.5: Main results – health parameters

	All		Boys		Girls		Boys - Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Diff.	S.E.
1st stage	0.56***	0.03	0.58***	0.04	0.55***	0.04	0.03	0.05
BMI	0.09	0.17	-0.13	0.23	0.35	0.25	-0.48	0.34
BMI (std)	0.02	0.05	-0.04	0.06	0.10	0.07	-0.13	0.09
Overweight (bin)	-0.01	0.01	-0.02	0.02	-0.01	0.02	-0.01	0.03
Obese (bin)	0.01	0.01	0.01	0.01	0.01	0.01	-0.01	0.01
Overweight or obese (bin)	-0.003	0.01	-0.01	0.02	0.004	0.02	-0.02	0.03
Weight in kg	0.28	0.55	0.10	0.72	0.46	0.75	-0.36	1.04
Weight in kg (std)	0.01	0.03	0.01	0.03	0.02	0.04	-0.02	0.05
Subjective health 1-5	-0.04	0.03	-0.05	0.04	-0.02	0.05	-0.03	0.06
Subjective health good (bin)	0.01	0.03	0.03	0.04	-0.01	0.04	0.03	0.05
Subjective health very good (bin)	-0.02	0.03	-0.04	0.03	-0.002	0.04	-0.03	0.05
Resting heart rate	-0.43	0.65	0.59	0.82	-1.44	0.95	2.02	1.26
Resting heart rate (std)	-0.04	0.05	0.05	0.07	-0.12	0.08	0.17	0.10
Height in cm	-0.02	0.46	0.52	0.63	-0.77	0.55	1.29	0.84
Height (std)	-0.001	0.03	0.03	0.04	-0.04	0.03	0.08	0.05
# of observations	4500		2342		2159			
# of clusters	485		435		425			
# of observations off support	320		85		234			
F-statistic of first stage	330.8***		260.9***		239.4***			
F-statistic for joint significane of LATE's	1.0		0.6		1.2			

Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level.
** p<0.1, ** p<0.05, *** p<0.01*

F: Heterogeneous effects

Table F.1 shows the results discussed in Section 6.2 for grades. As the other outcomes show mostly insignificant differences the other four tables are omitted.

Table F.1: Heterogeneity analysis for grades

Low income households (<2,500€ per month)						
	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.51***	0.04	0.51***	0.06	0.52***	0.05
German grade (std)	0.29***	0.08	0.30**	0.16	0.24**	0.11
Math grade (std)	0.27***	0.09	0.37***	0.13	0.14	0.13
Average grade (std)	0.32***	0.08	0.38***	0.14	0.22*	0.12
# of observations	1923		989		1000	
# of clusters	400		323		334	
# of observations off support	160		36		58	
F-statistic of first stage	162.4***		75.7***		130.9***	
F-statistic for joint significane of LATE's	5.0***		2.6**		1.6	

High income households (>2,500€ per month)						
	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.54***	0.04	0.57***	0.05	0.49***	0.05
German grade (std)	0.08	0.08	0.04	0.11	0.19*	0.11
Math grade (std)	0.04	0.08	0.07	0.11	0.04	0.13
Average grade (std)	0.07	0.08	0.06	0.10	0.13	0.12
# of observations	2016		982		954	
# of clusters	364		302		293	
# of observations off support	220		131		169	
F-statistic of first stage	186.4***		157.0***		114.5***	
F-statistic for joint significane of LATE's	0.3		0.1		0.9	

Differences between low and high income households						
	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	-0.03	0.06	-0.06	0.07	0.02	0.06
German grade (std)	0.20*	0.12	0.26	0.19	0.05	0.16
Math grade (std)	0.23*	0.12	0.30*	0.17	0.09	0.18
Average grade (std)	0.25**	0.12	0.33*	0.18	0.08	0.17

Notes: The three tables show the effects for low and high income households separately as well as the differences in the effects. Standard error are based on 4999 weighted bootstraps clustered at state-school type-class-wave level (p<0.1, ** p<0.05, *** p<0.01).*

G: Variable description

Table G.1: List of variables in the analysis and short description

Grades:	
German grade (std)	German grade (Larger means better grade)
Maths grade (std)	Math grade (Larger means better grade)
Average grade (std)	Average of German and math grade
Non-cognitive skills:	
Emotional symptoms index (std)	See Goodman (1997) for details
Conduct problems index (std)	
Hyperactivity index (std)	
Peer relations problems index (std)	
Asocial behaviour index (std)	
Total index (std)	
Emotional symptoms borderline or abnormal (bin)	
Emotional symptoms abnormal (bin)	
Conduct problems borderline or abnormal (bin)	
Conduct problems abnormal (bin)	
Hyperactivity borderline or abnormal (bin)	
Hyperactivity abnormal (bin)	
Peer problems borderline or abnormal (bin)	
Peer problems abnormal (bin)	
Asocial behaviour borderline or abnormal (bin)	
Asocial behaviour abnormal (bin)	
Total index borderline or abnormal (bin)	
Total index abnormal (bin)	
Extracurricular physical activity:	
Physical activity in club sports in minutes	Weekly minutes students participate in club sports
Physical activity in leisure sports in minutes	Weekly minutes students practise leisure sports
Physical activity in out of school in minutes	Sum of weekly minutes in club and leisure sports
Participation club sports (bin)	= 1 if students participate in club sports
# of days active per week (w/o PE)	Number of days students are at least 60 minutes moderate to vigorously active excluding PE
Compliance with WHO guideline (bin)	= 1 if children complies with the WHO guideline of daily 60 moderate to vigorous physical activity
Media consumption hrs/week	Hours of media consumption per week
# of club sports	Number of different club sports that students participate in
# of leisure sports	Number of different leisure sports that students practise
School-based physical activity:	
# of days with PE	Number of days per week with PE at school
# of voluntary lessons	Number of weekly lessons in voluntary PE

Table continues on next page >

Motor skills:	Exact description is found in Schmidt et al. (2016)
Push-ups (std)	Number of pushups students can do in 40 seconds
Side-steps (std)	How often students jump side to side repeatedly within 15 seconds
Static stand (std)	Standing on one leg on a 3 cm bar, how often does the second leg touch the ground
Standing long jump (std)	How far student jumps out of a static position
Reaction time (std)	Elapsed time after a color changes on a computer and students pushing a button
Balancing backwards (std)	Balancing backwards counting the steps until one foot touches the floor for the first time
Line tracking mistakes (std)	Number of mistakes tracing lines
Line tracking duration (std)	Mistake duration of tracing lines tracing lines
Line tracking duration (std)	Number of mistakes of tracing lines tracing lines
Inserting pins (std)	Time needed for sorting pins
Stand and reach (std)	Bend forward as far as students can, measure how far below or above the toes she can go
Health parameters:	
BMI	Body-Mass-Index (weight in kg / height in m)
BMI (std)	Body-Mass-Index (weight in kg / height in m)
Overweight (bin)	= 1 if 25 < BMI < 30
Obese (bin)	= 1 if BMI > 30
Overweight or obese (bin)	= 1 if BMI > 25
Weight in kg	Weight in kg
Weight in kg (std)	Weight in kg
Subjective health 1-5	Subjective health from 1 (very poor) to 5 (very good)
Subjective health good (bin)	Subjective health 4 or 5
Subjective health very good (bin)	Subjective health 5
Resting heart rate	Heart beats within 1 minute
Resting heart rate (std)	Heart beats within 1 minute
Height in cm	Height in cm
Height (std)	Height in cm
Control variables:	
Class level 1	= 1 if class level 1
Class level 2	= 1 if class level 2
Class level 3	= 1 if class level 3
Class level 4	= 1 if class level 4
Class level 5	= 1 if class level 5
Class level 6	= 1 if class level 6
Class level 7	= 1 if class level 7
Class level 8	= 1 if class level 8
Class level 9	= 1 if class level 9
Class level 10	= 1 if class level 10
Class level 11	= 1 if class level 11
Class level 12	= 1 if class level 12 or 13
Primary school	= 1 if student in primary school

Table continues on next page >

Control variables (continued):	
Basic / Intermediate school	= 1 if student in basic or intermediate school
Academic track	= 1 if student in academic track
Comprehensive school	= 1 if student in comprehensive school
HH income < 1,000€	= 1 if real household income (prices of 2010) smaller 1,000 €
HH income 1,000 - 1,500€	= 1 if real household income (prices of 2010) between 1,000 and 1,500 €
HH income 1,500 - 2,000€	= 1 if real household income (prices of 2010) between 1,500 and 2,000 €
HH income 2,000 - 2,500€	= 1 if real household income (prices of 2010) between 2,000 and 2,500 €
HH income 2,500 - 3,000€	= 1 if real household income (prices of 2010) between 2,500 and 3,000 €
HH income 3,000 - 4,000€	= 1 if real household income (prices of 2010) between 3,000 and 4,000 €
HH income 4,000 - 5,000€	= 1 if real household income (prices of 2010) between 4,000 and 5,000 €
HH income > 5,000€	= 1 if real household income (prices of 2010) larger 5,000 €
Low education HH	= 1 if parents education low
Middle education HH	= 1 if parents education middle
High education HH	= 1 if parents education high
Parents physically active	= 1 if both parents are physically active
Foreigner	= 1 if student is not German
No siblings	= 1 if student has no siblings
One sibling	= 1 if student has one sibling
Two siblings	= 1 if student has two siblings
Three or more siblings	= 1 if student has three or more siblings
Birthweight in kilogram	Birthweight in kilogram
Cohort of 1985 - 1990	= 1 if student born between 1985 and 1990
Cohort of 1991 - 1995	= 1 if student born between 1991 and 1995
Cohort of 1996 - 2000	= 1 if student born between 1996 and 2000
Cohort of 2001 - 2005	= 1 if student born between 2001 and 2005
Female	= 1 if student is female
East Germany	= 1 if student lives in east Germany
< 5,000 inhabitans	= 1 if hometown of student has less than 5,000 inhabitans
5,000 - 20,000 inhabitans	= 1 if hometown of student has between 5,000 and 10,000 inhabitans
20,000 - 100,000 inhabitans	= 1 if hometown of student has between 20,000 and 100,000 inhabitans
>100,000 inhabitans	= 1 if hometown of student has more than 100,000 inhabitans
Educ. exp. per student in 100€	Average expenditures per student in the *Länder* of the students

Notes: std: variable standardised to have zero mean and variance one, bin: binary indicator

Table G.2: Mean and standard deviation of all variables used in the analysis

	All		Boys		Girls	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Grades:						
German grade (std)	0.00	1.00	-0.23	1.01	0.23	0.94
Maths grade (std)	0.00	1.00	0.03	1.03	-0.03	0.97
Average grade (std)	0.00	1.00	-0.11	1.03	0.11	0.96
Non-cognitive skills:						
Emotional symptoms index (std)	0.00	1.00	-0.09	0.97	0.09	1.02
Conduct problems index (std)	0.00	1.00	0.10	1.04	-0.10	0.95
Hyperactivity index (std)	0.00	1.00	0.19	1.03	-0.19	0.93
Peer relations problems index (std)	0.00	1.00	0.08	1.05	-0.08	0.94
Asocial behaviour index (std)	0.00	1.00	0.17	1.05	-0.17	0.92
Total index (std)	0.00	1.00	0.11	1.04	-0.11	0.95
Emotional symptoms borderline or abnormal (bin)	0.16	0.37	0.14	0.35	0.18	0.39
Emotional symptoms abnormal (bin)	0.09	0.29	0.08	0.27	0.10	0.30
Conduct problems borderline or abnormal (bin)	0.29	0.45	0.33	0.47	0.24	0.43
Conduct problems abnormal (bin)	0.13	0.34	0.16	0.36	0.11	0.31
Hyperactivity borderline or abnormal (bin)	0.12	0.33	0.17	0.37	0.08	0.28
Hyperactivity abnormal (bin)	0.07	0.26	0.10	0.30	0.05	0.21
Peer relations problems borderline or abnormal (bin)	0.18	0.39	0.21	0.41	0.15	0.36
Peer relations problems abnormal (bin)	0.09	0.29	0.11	0.31	0.08	0.27
Asocial behaviour borderline or abnormal (bin)	0.08	0.26	0.10	0.30	0.05	0.22
Asocial behaviour abnormal (bin)	0.02	0.15	0.03	0.17	0.02	0.12
Total index borderline or abnormal (bin)	0.13	0.33	0.15	0.36	0.10	0.30
Total index abnormal (bin)	0.06	0.24	0.07	0.26	0.05	0.21
Extracurricular physical activity:						
Minutes club sports	61.49	65.58	66.33	65.60	56.67	65.22
Minutes leisure sports	73.87	124.69	87.10	142.39	60.70	102.50
Minutes active out of school	135.35	139.40	153.43	153.53	117.37	121.14
Participation club sports (bin)	0.63	0.48	0.69	0.46	0.57	0.50
# of days active per week	3.86	1.82	4.07	1.80	3.65	1.83
Compliance with WHO guideline (bin)	0.14	0.32	0.16	0.33	0.12	0.30
Media consumption hrs/week	17.39	14.43	19.77	16.04	15.03	12.18
# of club sports	0.84	0.79	0.90	0.76	0.78	0.82
# of leisure sports	1.00	1.13	1.02	1.13	0.97	1.13
School-based physical activity:						
# of days with PE	1.82	0.78	1.84	0.79	1.80	0.77
# of voluntary lessons	0.38	1.02	0.42	1.08	0.33	0.94
Motor skills:						
Pushups (std)	0.00	1.00	0.10	1.06	-0.11	0.93
Jumping side to side (std)	0.00	1.00	-0.01	1.06	0.01	0.93

Table continues on next page >

	All		Boys		Girls	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Motor skills (continued):						
Single leg stance (std)	0.00	1.00	0.11	1.07	-0.11	0.92
Standing long jum (std)	0.00	1.00	0.28	1.10	-0.29	0.78
Reaction test (std)	0.00	1.00	-0.06	0.94	0.06	1.05
Backwards balancing (std)	0.00	1.00	-0.12	1.02	0.12	0.96
Tracing lines mistakes (std)	0.00	1.00	0.20	1.04	-0.20	0.91
Tracing lines mistake duration (std)	0.00	1.00	0.15	1.12	-0.15	0.83
Tracing lines duration (std)	0.00	1.00	-0.04	0.97	0.04	1.03
Sorting pens duration (std)	0.00	1.00	0.13	1.03	-0.13	0.96
Forward bend (std)	0.00	1.00	-0.30	0.92	0.31	0.99
Health & Fitness:						
BMI	19.11	3.67	19.11	3.69	19.11	3.65
BMI (std)	0.00	1.00	0.00	1.00	0.00	1.00
Overweight (bin)	0.06	0.24	0.06	0.25	0.05	0.22
Obese (bin)	0.01	0.11	0.01	0.17	0.01	0.11
Overweight or obese (bin)	0.07	0.26	0.08	0.27	0.06	0.25
Weight in kg	46.21	21.48	47.19	17.72	45.2	24.68
Weight in kg (std)	0.00	1.00	0.05	0.82	-0.05	1.14
Subjective health 1-5	4.42	0.60	4.42	0.60	4.43	0.59
Subjective health good (bin)	0.47	0.50	0.48	0.50	0.47	0.49
Subjective health very good (bin)	0.48	0.50	0.47	0.49	0.48	0.50
Resting heart rate	80.72	12.18	79.23	11.99	82.2	12.18
Resting heart rate (std)	0.00	1.00	-0.12	0.98	0.12	1.00
Height in cm	152.47	17.25	154.14	18.51	150.78	15.70
Height (std)	0.00	1.00	0.10	1.07	-0.10	0.91
Control variables:						
Class level 1	0.08	0.27	0.08	0.27	0.07	0.25
Class level 2	0.09	0.29	0.09	0.29	0.09	0.28
Class level 3	0.10	0.30	0.09	0.28	0.10	0.30
Class level 4	0.09	0.29	0.09	0.29	0.09	0.28
Class level 5	0.10	0.30	0.10	0.30	0.10	0.29
Class level 6	0.10	0.31	0.10	0.30	0.11	0.30
Class level 7	0.10	0.31	0.11	0.31	0.10	0.29
Class level 8	0.10	0.30	0.10	0.30	0.09	0.29
Class level 9	0.10	0.30	0.09	0.29	0.10	0.30
Class level 10	0.08	0.27	0.08	0.27	0.08	0.27
Class level 11	0.04	0.20	0.04	0.19	0.05	0.20
Class level 12	0.02	0.13	0.01	0.11	0.02	0.14
Primary school	0.37	0.48	0.37	0.48	0.37	0.48
Basic / Intermediate school	0.27	0.44	0.29	0.45	0.24	0.42
Academic track	0.31	0.46	0.28	0.45	0.35	0.47
Comprehensive school	0.05	0.23	0.06	0.23	0.05	0.21

Table continues on next page >

	All		Boys		Girls	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Control variables (continued):						
HH income < 1,000€	0.11	0.32	0.11	0.31	0.12	0.32
HH income 1,000 - 1,500€	0.08	0.27	0.08	0.27	0.07	0.26
HH income 1,500 - 2,000€	0.12	0.32	0.11	0.32	0.12	0.32
HH income 2,000 - 2,500€	0.18	0.39	0.18	0.39	0.18	0.38
HH income 2,500 - 3,000€	0.18	0.38	0.18	0.39	0.17	0.37
HH income 3,000 - 4,000€	0.19	0.39	0.19	0.39	0.19	0.39
HH income 4,000 - 5,000€	0.10	0.30	0.09	0.29	0.11	0.30
HH income > 5,000€	0.05	0.23	0.06	0.24	0.05	0.21
Low education HH	0.11	0.32	0.11	0.31	0.12	0.31
Middle education HH	0.57	0.50	0.58	0.49	0.57	0.49
High education HH	0.32	0.47	0.32	0.46	0.32	0.46
Parents physically active	0.26	0.44	0.25	0.44	0.26	0.43
Foreigner	0.03	0.17	0.03	0.16	0.03	0.17
No siblings	0.14	0.35	0.14	0.38	0.14	0.34
One sibling	0.50	0.50	0.50	0.50	0.49	0.50
Two siblings	0.25	0.43	0.24	0.45	0.26	0.43
Three or more siblings	0.12	0.32	0.12	0.32	0.12	0.32
Birthweight in kilogram	3.38	0.55	3.46	0.56	3.31	0.54
Cohort of 1985 - 1990	0.13	0.34	0.13	0.33	0.13	0.33
Cohort of 1991 - 1995	0.33	0.47	0.34	0.48	0.33	0.46
Cohort of 1996 - 2000	0.38	0.49	0.38	0.48	0.38	0.49
Cohort of 2001 - 2005	0.16	0.36	0.15	0.36	0.16	0.36
Female	0.50	0.50	0.00	0.00	1.00	0.00
East Germany	0.32	0.47	0.32	0.47	0.33	0.47
< 5,000 inhabitans	0.25	0.43	0.26	0.44	0.25	0.43
5,000 - 20,000 inhabitans	0.33	0.47	0.32	0.47	0.33	0.47
20,000 - 100,000 inhabitans	0.30	0.46	0.29	0.45	0.30	0.46
>100,000 inhabitans	0.13	0.33	0.13	0.33	0.13	0.34
Educ. exp. per student in 100€	57.10	7.09	56.98	7.10	57.23	7.07
# of Observations	5,243		2,704		2,719	

Notes: std: variable standardised to have zero mean and variance one, bin: binary indicator

H: Sensitivity analyses

H.1: Controlling for pre-school differences in outcomes

Table H.1.1: Sensitivity pre-school differences - Non-cognitive skills

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Emotional symptoms index (std)	-0.04	0.05	-0.002	0.06	-0.07***	0.08
Conduct problems index (std)	0.13***	0.05	0.16***	0.06	0.09	0.07
Hyperactivity index (std)	-0.05	0.04	-0.10	0.07	0.004	0.06
Peer relations problems index (std)	0.08	0.05	0.24***	0.07	-0.07	0.07
Asocial behaviour index (std)	-0.01	0.05	-0.04	0.07	0.02	0.07
Total index (std)	0.02	0.04	0.06	0.06	-0.01	0.07
Emotional symptoms borderline or abnormal (bin)	0.003	0.02	0.02	0.02	-0.03	0.03
Emotional symptoms abnormal (bin)	-0.04**	0.02	-0.01	0.02	-0.06	0.03
Conduct problems borderline or abnormal (bin)	0.05**	0.02	0.06*	0.03	0.04**	0.03
Conduct problems abnormal (bin)	0.02	0.02	0.04*	0.02	0.003	0.02
Hyperactivity borderline or abnormal (bin)	0.02	0.02	0.02	0.03	0.02	0.02
Hyperactivity abnormal (bin)	0.01	0.01	0.02	0.02	-0.01	0.02
Peer relations problems borderline or abnormal (bin)	0.01	0.02	0.08***	0.03	-0.05	0.02
Peer relations problems abnormal (bin)	0.002	0.01	0.05***	0.02	-0.05*	0.02
Asocial behaviour borderline or abnormal (bin)	-0.01	0.01	0.001	0.02	-0.02	0.02
Asocial behaviour abnormal (bin)	-0.01	0.01	-0.01	0.01	-0.01	0.01
Total index borderline or abnormal (bin)	0.02	0.02	0.06***	0.02	-0.02*	0.02
Total index abnormal (bin)	0.03**	0.01	0.04**	0.02	0.02	0.02
# of observations						
# of clusters	*Slightly different because each specification has ist own common support procedure. However, very close to the numbers of the main specification.*					
# of observations off support						
F-statistic of first stage						
F-statistic for joint significane of LATE's						

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, educational spending per students, and state-average pre-school levels of the respective outcomes. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.1.2: Sensitivity pre-school differences - Motor skills

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Push-ups (std)	*not measured for pre-school children*					
Side-steps (std)	-0.03	0.05	-0.03	0.06	-0.04	0.07
Static stand (std)	0.02	0.05	0.03	0.08	0.002	0.07
Standing long jump (std)	0.02	0.04	0.03	0.05	0.02	0.06
Reaction time (std)	0.04	0.05	0.03	0.05	0.06	0.07
Balancing backwards (std)	0.11*	0.06	0.07	0.08	0.13*	0.08
Line tracking mistakes (std)	-0.04	0.05	-0.05	0.07	0.01	0.07
Line tracking mistake duration (std)	-0.06	0.08	-0.04	0.11	-0.01	0.07
Line tracking duration (std)	0.13**	0.07	0.11	0.09	0.18**	0.09
Inserting pins duration (std)	-0.08*	0.05	-0.07	0.06	-0.10	0.06
Stand and reach (std)	0.14***	0.05	0.05	0.07	0.25***	0.08
# of observations						
# of clusters	*Slightly different because each specification has ist own common support procedure. However, very close to the numbers of the main specification.*					
# of observations off support						
F-statistic of first stage						
F-statistic for joint significane of LATE's						

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, educational spending per students, and state-average pre-school levels of the respective outcomes. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.1.3: Sensitivity pre-school differences - Physical activity

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Extracurricular:						
Physical activity in club sports in minutes	-0.91	7.20	2.17	5.69	-5.10	6.33
Physical activity in leisure sports in minutes	-2.21	6.58	1.50	9.68	-5.34	8.39
Physical activity in out of school in minutes	-2.74	6.93	1.31	10.67	-5.29	9.23
Participation club sports (bin)	-0.006	0.03	-0.04	0.04	0.02	0.05
# of days active per week	0.15*	0.10	0.11	0.14	0.26**	0.13
Compliance with WHO guideline (bin)	-0.01	0.01	-0.02	0.02	-0.01	0.02
Media consumption hrs/week	0.55	0.74	1.22	1.07	0.54	0.85
# of club sports	0.01	0.05	0.02	0.06	-0.001	0.07
# of leisure sports	0.14**	0.06	0.08	0.08	0.22***	0.09
# of observations						
# of clusters	*Slightly different because each specification has ist own common support procedure. However, very close to the numbers of the main specification.*					
# of observations off support						
F-statistic of first stage						
F-statistic for joint significane of LATE's						

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, educational spending per students, and state-average pre-school levels of the respective outcomes. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.1.4: Sensitivity pre-school differences - Health parameters

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
BMI	0.08	0.17	-0.12	0.23	0.32	0.25
BMI (std)	0.02	0.05	-0.03	0.07	0.09	0.07
Overweight (bin)	-0.01	0.02	-0.02	0.02	-0.002	0.08
Overweight or obese (bin)	-0.01	0.02	-0.02	0.12	0.004	0.02
Weight in kg	0.55	0.56	0.42	0.75	0.75	0.72
Weight in kg (std)	0.03	0.03	0.02	0.03	0.04	0.03
Subjective health 1-5	-0.04	0.03	-0.06	0.04	-0.01	0.05
Subjective health good (bin)	0.03	0.03	0.03	0.04	0.03	0.04
Subjective health very good (bin)	-0.03	0.03	-0.04	0.03	-0.01	0.04
Resting heart rate	-0.38	0.64	0.52	0.85	-1.17	0.92
Resting heart rate (std)	-0.03	0.05	0.04	0.07	-0.10	0.08
Height in cm	0.43	0.47	1.01	0.61	-0.38	0.60
Height (std)	0.03	0.02	0.06	0.04	-0.02	0.03
# of observations						
# of clusters	*Slightly different because each specification has ist own common support procedure. However, very close to the numbers of the main specification.*					
# of observations off support						
F-statistic of first stage						
F-statistic for joint significane of LATE's						

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, educational spending per students, and state-average pre-school levels of the respective outcomes. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

H.2: Excluding outlier in PE lessons

Table H.2.1: Sensitivity outlier - Grades

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.43***	0.03	0.43***	0.04	0.43	0.04
German grade (std)	0.24***	0.08	0.15	0.12	0.29	0.10
Math grade (std)	0.17**	0.08	0.17	0.12	0.14	0.11
Average grade (std)	0.24***	0.08	0.18	0.12	0.24	0.10
# of observations	3935		2011		1931	
# of clusters	440		395		382	
# of observations off support	272		63		202	
F-statistic of first stage	203.4***		143.0***		149.3***	
F-statistic for joint significane of LATE's	3.3**		0.8		3.0**	

*Notes: This table shows the weighted LATE estimates for students of the 3rd to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.2.2: Sensitivity outlier - Non-cognitive skills

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.48***	0.03	0.48***	0.03	0.49***	0.03
Emotional symptoms index (std)	-0.06	0.06	-0.04	0.08	-0.08	0.09
Conduct problems index (std)	0.14**	0.05	0.19**	0.08	0.09	0.08
Hyperactivity index (std)	-0.04	0.05	-0.09	0.09	0.02	0.07
Peer relations problems index (std)	0.11*	0.05	0.27***	0.09	-0.04	0.08
Asocial behaviour index (std)	-0.07	0.06	-0.15	0.10	-0.02	0.08
Total index (std)	0.04	0.05	0.09	0.08	-0.005	0.08
Emotional symptoms borderline or abnormal (bin)	-0.03	0.02	-0.002	0.03	-0.05	0.03
Emotional symptoms abnormal (bin)	-0.04**	0.01	-0.02	0.02	-0.06**	0.03
Conduct problems borderline or abnormal (bin)	0.04*	0.02	0.06	0.04	0.03	0.04
Conduct problems abnormal (bin)	0.02	0.02	0.04	0.03	0.003	0.03
Hyperactivity borderline or abnormal (bin)	0.02	0.01	0.02	0.03	0.02	0.02
Hyperactivity abnormal (bin)	0.01	0.02	0.02	0.02	-0.004	0.02
Peer relations problems borderline or abnormal (bin)	0.02	0.02	0.09**	0.04	-0.05*	0.03
Peer relations problems abnormal (bin)	0.01	0.02	0.06**	0.02	-0.03	0.02
Asocial behaviour borderline or abnormal (bin)	-0.01	0.02	0.003	0.03	-0.03	0.02
Asocial behaviour abnormal (bin)	-0.01	0.01	-0.01	0.02	-0.02*	0.01
Total index borderline or abnormal (bin)	0.03	0.02	0.08***	0.03	-0.01	0.03
Total index abnormal (bin)	0.02*	0.01	0.04**	0.02	0.01	0.02
# of observations	4909		2509		2413	
# of clusters	491		440		430	
# of observations off support	366		108		245	
F-statistic of first stage	304.5***		226.2***		230.1***	
F-statistic for joint significane of LATE's	2.1***		2.6***		1.4	

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.2.3: Sensitivity outlier - Motor skills

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.47***	0.03	0.45***	0.04	0.48***	0.03
Push-ups (std)	-0.13**	0.06	-0.13	0.10	-0.16*	0.08
Side-steps (std)	0.09	0.05	0.09	0.07	0.08	0.07
Static stand (std)	0.05	0.06	0.06	0.11	0.06	0.08
Standing long jump (std)	0.02	0.05	0.02	0.06	-0.001	0.07
Reaction time (std)	0.07	0.06	0.07	0.07	0.06	0.08
Balancing backwards (std)	0.12**	0.06	0.07	0.10	0.17**	0.08
Line tracking mistakes (std)	-0.04	0.06	-0.06	0.10	0.02	0.07
Line tracking mistake duration (std)	-0.04	0.07	-0.03	0.12	-0.04	0.07
Line tracking duration (std)	0.15*	0.08	0.12	0.12	0.16	0.10
Inserting pins duration (std)	-0.10*	0.06	-0.09	0.07	-0.07	0.07
Stand and reach (std)	0.18***	0.06	0.05	0.09	0.30***	0.09
# of observations	4188		2159		2010	
# of clusters	479		431		416	
# of observations off support	300		85		234	
F-statistic of first stage	245.4***		163.7***		207.1***	
F-statistic for joint significane of LATE's	3.1***		1.2		2.3***	

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.2.4: Sensitivity outlier - Physical activity

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.48***	0.03	0.48***	0.03	0.50***	0.03
School based:						
# of days with PE	0.92***	0.04	0.91***	0.05	0.92***	0.05
# of voluntary lessons	-0.07	0.07	-0.08	0.11	-0.05	0.08
F-statistic for joint significane of LATE's	225.6***		155.8***		152.5***	
Extracurricular:						
Physical activity in club sports in minutes	-4.87	4.04	-2.16	5.73	-5.76	5.36
Physical activity in leisure sports in minutes	4.94	7.72	12.62	12.22	-1.49	9.33
Physical activity in out of school in minutes	0.07	8.96	10.47	13.71	-7.25	10.50
Participation club sports (bin)	-0.05*	0.03	-0.06	0.04	-0.04	0.04
# of days active per week	0.20*	0.12	0.11	0.17	0.25*	0.14
Compliance with WHO guideline (bin)	-0.01	0.02	-0.02	0.03	0.003	0.02
Media consumption hrs/week	0.86	0.84	1.27	1.27	0.68	0.95
# of club sports	-0.06	0.05	-0.01	0.07	-0.09	0.07
# of leisure sports	0.21**	0.08	0.16	0.11	0.26**	0.10
# of observations	4592		2331		2342	
# of clusters	483		429		424	
# of observations off support	339		108		150	
F-statistic of first stage	295.5***		216.6***		244.1***	
F-statistic for joint significane of LATE's	2.7***		1.4		2.0**	

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.2.5: Sensitivity outlier - Health parameters

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.47***	0.03	0.45***	0.04	0.48***	0.03
BMI	0.12	0.20	-0.12	0.30	0.41	0.28
BMI (std)	0.03	0.05	-0.03	0.08	0.11	0.08
Overweight (bin)	-0.02	0.02	-0.03	0.03	-0.01	0.02
Obese (bin)	0.01	0.01	0.01	0.01	0.02	0.01
Overweight or obese (bin)	-0.01	0.02	-0.02	0.03	0.01	0.02
Weight in kg	0.52	0.65	0.34	0.94	0.69	0.85
Weight in kg (std)	0.02	0.03	0.02	0.04	0.03	0.04
Subjective health 1-5	-0.06	0.04	-0.08	0.05	-0.02	0.05
Subjective health good (bin)	0.02	0.03	0.04	0.05	-0.01	0.05
Subjective health very good (bin)	-0.03	0.03	-0.06	0.04	-0.003	0.04
Resting heart rate	-0.50	0.80	1.08	1.12	-1.79	1.12
Resting heart rate (std)	-0.04	0.07	0.09	0.09	-0.15	0.09
Height in cm	0.25	0.56	0.90	0.82	-0.64	0.63
Height (std)	0.01	0.03	0.05	0.05	-0.04	0.04
# of observations	4365		2276		2098	
# of clusters	481		434		420	
# of observations off support	328		75		244	
F-statistic of first stage	257.8***		171.2***		205.7***	
F-statistic for joint significane of LATE's	1.1		0.7		1.0	

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

H.3: Consider sampling weights

Table H.3.1: Sensitivity sampling weights - Grades

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.48***	0.04	0.57***	0.05	0.42***	0.05
German grade (std)	0.30***	0.10	0.23*	0.12	0.36***	0.13
Math grade (std)	0.18**	0.09	0.24**	0.11	0.04	0.14
Average grade (std)	0.27***	0.09	0.27**	0.11	0.22*	0.13
# of observations	3801		1950		1875	
# of clusters	431		385		376	
# of observations off support	518		188		306	
F-statistic of first stage	154.7***		124.7***		86.1***	
F-statistic for joint significane of LATE's	3.4**		1.9		2.7**	

*Notes: This table shows the weighted LATE estimates for students of the 3rd to 12th grade. Inverse Probability Tilting with sampling weights is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.3.2: Sensitivity sampling weights - Non-cognitive skills

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.55***	0.04	0.63***	0.05	0.47***	0.04
Emotional symptoms index (std)	-0.07	0.08	-0.09	0.09	-0.04	0.13
Conduct problems index (std)	0.11*	0.07	0.10	0.09	0.12	0.11
Hyperactivity index (std)	-0.16**	0.06	-0.24***	0.08	-0.04	0.11
Peer relations problems index (std)	0.14**	0.07	0.25***	0.09	-0.004	0.11
Asocial behaviour index (std)	-0.14**	0.07	-0.20**	0.09	-0.05	0.10
Total index (std)	-0.02	0.07	-0.03	0.08	0.002	0.12
Emotional symptoms borderline or abnormal (bin)	0.002	0.03	0.03	0.04	-0.02	0.05
Emotional symptoms abnormal (bin)	-0.04	0.02	-0.03	0.03	-0.04	0.04
Conduct problems borderline or abnormal (bin)	0.03	0.03	0.02	0.04	0.04	0.05
Conduct problems abnormal (bin)	0.02	0.03	0.02	0.03	0.03	0.04
Hyperactivity borderline or abnormal (bin)	-0.04	0.03	-0.08**	0.04	0.02	0.02
Hyperactivity abnormal (bin)	-0.03	0.02	-0.03	0.03	-0.03	0.02
Peer relations problems borderline or abnormal (bin)	0.02	0.03	0.06*	0.03	-0.04	0.04
Peer relations problems abnormal (bin)	0.02	0.02	0.04*	0.03	-0.02	0.03
Asocial behaviour borderline or abnormal (bin)	-0.03	0.02	-0.03	0.03	-0.02	0.02
Asocial behaviour abnormal (bin)	-0.01	0.01	-0.02	0.02	-0.01	0.02
Total index borderline or abnormal (bin)	0.004	0.03	0.02	0.03	-0.01	0.04
Total index abnormal (bin)	0.01	0.02	0.02	0.02	-0.01	0.03
# of observations	4874		2373		2456	
# of clusters	485		424		436	
# of observations off support	545		328		262	
F-statistic of first stage	236.3***		182.2***		142.3***	
F-statistic for joint significane of LATE's	2.9***		2.7***		1.0	

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting with sampling weights is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.3.3: Sensitivity sampling weights - Motor skills

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.54***	0.04	0.61***	0.05	0.47***	0.04
Push-ups (std)	-0.09	0.07	-0.08	0.10	-0.10	0.11
Side-steps (std)	0.07	0.05	0.04	0.07	0.08	0.08
Static stand (std)	0.06	0.08	-0.01	0.10	0.17*	0.10
Standing long jump (std)	-0.03	0.05	0.04	0.06	-0.06	0.08
Reaction time (std)	0.12**	0.06	0.11*	0.06	0.11	0.09
Balancing backwards (std)	0.06	0.07	0.09	0.09	0.06	0.10
Line tracking mistakes (std)	-0.05	0.07	-0.10	0.08	0.07	0.09
Line tracking mistake duration (std)	-0.05	0.07	-0.08	0.08	0.01	0.07
Line tracking duration (std)	0.11	0.08	0.12	0.10	0.10	0.12
Inserting pins duration (std)	-0.04	0.05	-0.05	0.06	-0.02	0.09
Stand and reach (std)	0.18**	0.07	0.14*	0.08	0.26**	0.11
# of observations	4175		2092		2115	
# of clusters	475		422		423	
# of observations off support	430		226		172	
F-statistic of first stage	208.6***		166.8***		123.5***	
F-statistic for joint significane of LATE's	2.2**		1.5		1.5	

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting with sampling weights is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.3.4: Sensitivity sampling weights - Physical activity

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.56***	0.04	0.60***	0.05	0.49***	0.04
School based:						
# of days with PE	0.74***	0.04	0.71***	0.06	0.81***	0.06
# of voluntary lessons	0.001	0.07	-0.01	0.10	-0.01	0.08
F-statistic for joint significane of LATE's	142.0***		85.0***		105.7***	
Extracurricular:						
Physical activity in club sports in minutes	-3.59	4.38	0.22	5.96	-5.33	6.39
Physical activity in leisure sports in minutes	-8.67	9.88	-3.69	13.78	-17.41	15.45
Physical activity in out of school in minutes	-12.27	10.68	-3.46	15.57	-22.74	15.38
Participation club sports (bin)	-0.02	0.03	-0.05	0.04	0.02	0.05
# of days active per week	0.16	0.13	0.13	0.17	0.13	0.20
Compliance with WHO guideline (bin)	-0.01	0.05	-0.03	0.03	-0.01	0.03
Media consumption hrs/week	-0.17	1.06	0.72	1.32	0.25	1.47
# of club sports	-0.03	0.06	-0.01	0.07	-0.02	0.08
# of leisure sports	0.11	0.08	0.12	0.10	0.06	0.13
# of observations	4692		2200		2364	
# of clusters	483		414		429	
# of observations off support	376		317		187	
F-statistic of first stage	245.8***		163.4***		160.4***	
F-statistic for joint significane of LATE's	1.5		1.7*		1.1	

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting with sampling weights is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.3.5: Sensitivity sampling weights - Health parameters

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.53***	0.04	0.58***	0.05	0.48***	0.04
BMI	0.23	0.25	0.07	0.29	0.70*	0.42
BMI (std)	0.08	0.07	0.02	0.08	0.19*	0.12
Overweight (bin)	0.02	0.02	0.03	0.02	0.01	0.03
Obese (bin)	0.01	0.01	0.001	0.01	0.01	0.02
Overweight or obese (bin)	0.02	0.02	0.03	0.03	0.02	0.03
Weight in kg	0.52	0.83	0.48	0.97	1.23	1.26
Weight in kg (std)	0.02	0.04	0.02	0.05	0.06	0.06
Subjective health 1-5	-0.05	0.04	-0.03	0.05	-0.06	0.07
Subjective health good (bin)	0.03	0.04	0.01	0.05	0.04	0.05
Subjective health very good (bin)	-0.03	0.04	-0.02	0.04	-0.04	0.05
Resting heart rate	-0.68	0.82	-0.44	1.01	-1.38	1.27
Resting heart rate (std)	-0.06	0.07	-0.04	0.08	-0.11	0.10
Height in cm	-0.41	0.67	0.37	0.84	-0.84	0.78
Height (std)	-0.02	0.04	0.02	0.05	-0.05	0.05
# of observations	4325		2158		2190	
# of clusters	477		422		427	
# of observations off support	495		269		203	
F-statistic of first stage	214.7***		165.3***		128.4***	
F-statistic for joint significane of LATE's	0.8		0.3		1.0	

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting with sampling weights is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

H.4: Consider only grade three and older

Table H.4.1: Sensitivity grade three and older - Grades

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.52***	0.03	0.55***	0.04	0.50***	0.04
German grade (std)	0.21***	0.06	0.13	0.09	0.25***	0.08
Math grade (std)	0.16***	0.06	0.15*	0.09	0.12	0.09
Average grade (std)	0.21***	0.06	0.17*	0.09	0.21**	0.09
# of observations	4035		2055		1967	
# of clusters	443		392		388	
# of observations off support	284		83		214	
F-statistic of first stage	236.7***		201.6***		178.6***	
F-statistic for joint significane of LATE's	4.4***		1.1		3.0**	

*Notes: This table shows the weighted LATE estimates for students of the 3rd to 12th grade. Inverse Probability Tilting with sampling weights is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.4.2: Sensitivity grade three and older - Non-cognitive skills

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.53***	0.03	0.56***	0.04	0.50***	0.04
Emotional symptoms index (std)	-0.01	0.05	-0.01	0.08	0.01	0.09
Conduct problems index (std)	0.13**	0.06	0.18**	0.07	0.09	0.08
Hyperactivity index (std)	-0.03	0.05	-0.06	0.08	0.03	0.07
Peer relations problems index (std)	0.12**	0.05	0.28***	0.08	-0.01	0.08
Asocial behaviour index (std)	-0.02	0.05	-0.03	0.09	-0.01	0.08
Total index (std)	0.06	0.05	0.10	0.07	0.04	0.08
Emotional symptoms borderline or abnormal (bin)	-0.01	0.02	0.01	0.03	-0.03	0.03
Emotional symptoms abnormal (bin)	-0.03	0.02	-0.01	0.02	-0.05*	0.03
Conduct problems borderline or abnormal (bin)	0.05*	0.02	0.06*	0.04	0.04	0.04
Conduct problems abnormal (bin)	0.04*	0.02	0.06**	0.03	0.01	0.03
Hyperactivity borderline or abnormal (bin)	0.01	0.02	0.004	0.03	0.03	0.02
Hyperactivity abnormal (bin)	0.01	0.01	0.02	0.02	0.001	0.07
Peer relations problems borderline or abnormal (bin)	0.02	0.02	0.08***	0.03	-0.04	0.03
Peer relations problems abnormal (bin)	0.01	0.02	0.04	0.02	-0.02	0.02
Asocial behaviour borderline or abnormal (bin)	-0.01	0.01	0.03	0.02	-0.04**	0.02
Asocial behaviour abnormal (bin)	-0.02**	0.01	-0.01	0.01	-0.03**	0.01
Total index borderline or abnormal (bin)	0.03*	0.02	0.08***	0.02	-0.01	0.03
Total index abnormal (bin)	0.03**	0.01	0.04**	0.02	0.02	0.02
# of observations	4231		2133		2054	
# of clusters	449		395		391	
# of observations off support	279		95		228	
F-statistic of first stage	250.3***		212.6***		183.6***	
F-statistic for joint significane of LATE's	2.0***		2.6***		1.5*	

*Notes: This table shows the weighted LATE estimates for students of the 3rd to 12th grade. Inverse Probability Tilting with sampling weights is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.4.3: Sensitivity grade three and older - Motor skills

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.51***	0.03	0.53***	0.04	0.48***	0.04
Push-ups (std)	-0.10	0.06	-0.09	0.09	-0.17*	0.09
Side-steps (std)	0.12**	0.05	0.07	0.07	0.13*	0.08
Static stand (std)	0.03	0.06	0.06	0.09	0.01	0.08
Standing long jump (std)	0.04	0.05	0.06	0.06	-0.02	0.07
Reaction time (std)	0.05	0.05	0.03	0.06	0.08	0.09
Balancing backwards (std)	0.11*	0.06	0.004	0.09	0.23***	0.08
Line tracking mistakes (std)	-0.07	0.06	-0.01	0.09	0.002	0.07
Line tracking mistake duration (std)	-0.10	0.07	-0.11	0.09	-0.07	0.08
Line tracking duration (std)	0.12	0.08	0.08	0.10	0.13	0.10
Inserting pins duration (std)	-0.12**	0.05	-0.08	0.06	-0.12*	0.07
Stand and reach (std)	0.22***	0.07	0.12	0.09	0.33***	0.10
# of observations	3627		1871		1720	
# of clusters	437		388		376	
# of observations off support	223		61		198	
F-statistic of first stage	220.1***		164.0***		147.8***	
F-statistic for joint significane of LATE's	3.2***		1.1		2.5***	

*Notes: This table shows the weighted LATE estimates for students of the 3rd to 12th grade. Inverse Probability Tilting with sampling weights is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.4.4: Sensitivity grade three and older - Physical activity

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.54***	0.03	0.55***	0.04	0.52***	0.04
School based:						
# of days with PE	0.85***	0.04	0.84***	0.05	0.85***	0.06
# of voluntary lessons	-0.07	0.07	-0.06	0.11	-0.08	0.09
F-statistic for joint significane of LATE's	192.4***		157.7***		116.4***	
Extracurricular:						
Physical activity in club sports in minutes	0.45	3.74	1.96	5.26	1.10	5.54
Physical activity in leisure sports in minutes	6.19	8.00	12.44	11.41	-0.05	10.39
Physical activity in out of school in minutes	6.64	9.02	14.40	12.97	1.05	11.56
Participation club sports (bin)	-0.02	0.03	-0.03	0.04	-0.004	0.04
# of days active per week	0.24**	0.12	0.24	0.17	0.27*	0.15
Compliance with WHO guideline (bin)	-0.01	0.02	-0.03	0.03	0.01	0.02
Media consumption hrs/week	0.62	0.87	1.02	1.29	0.30	1.10
# of club sports	0.004	0.04	0.04	0.06	-0.01	0.07
# of leisure sports	0.16***	0.08	0.10	0.09	0.23**	0.11
# of observations	3946		1995		1983	
# of clusters	439		383		387	
# of observations off support	286		94		160	
F-statistic of first stage	249.9***		193.9***		187.2***	
F-statistic for joint significane of LATE's	2.7***		2.8***		1.3	

*Notes: This table shows the weighted LATE estimates for students of the 3rd to 12th grade. Inverse Probability Tilting with sampling weights is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.4.5: Sensitivity grade three and older - Health parameters

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.51***	0.04	0.54***	0.04	0.49***	0.04
BMI	0.05	0.22	-0.25	0.29	0.41	0.34
BMI (std)	0.01	0.06	-0.07	0.08	0.11	0.09
Overweight (bin)	-0.02	0.02	-0.02	0.02	-0.01	0.03
Obese (bin)	0.01	0.01	0.01	0.01	0.02	0.01
Overweight or obese (bin)	-0.004	0.02	-0.02	0.03	0.01	0.03
Weight in kg	0.29	0.70	-0.19	0.91	0.80	1.00
Weight in kg (std)	0.01	0.03	-0.01	0.04	0.04	0.05
Subjective health 1-5	-0.06	0.04	-0.06	0.05	-0.06	0.06
Subjective health good (bin)	0.003	0.03	0.01	0.04	0.01	0.05
Subjective health very good (bin)	-0.03	0.03	-0.03	0.04	-0.03	0.05
Resting heart rate	-0.47	0.79	0.29	0.97	-1.12	1.17
Resting heart rate (std)	-0.04	0.06	0.02	0.08	-0.09	0.10
Height in cm	0.10	0.57	0.49	0.74	-0.58	0.67
Height (std)	0.01	0.03	0.03	0.04	-0.03	0.04
# of observations	3810		1966		1803	
# of clusters	440		390		382	
# of observations off support	223		57		207	
F-statistic of first stage	218.9***		171.2***		146.9***	
F-statistic for joint significane of LATE's	1.0		0.6		0.8	

*Notes: This table shows the weighted LATE estimates for students of the 3rd to 12th grade. Inverse Probability Tilting with sampling weights is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

H.5: Balanced sample

Table H.5.1: Sensitivity balanced sample - Grades

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.51***	0.04	0.52***	0.04	0.50***	0.04
German grade (std)	0.23***	0.07	0.16	0.10	0.25***	0.09
Math grade (std)	0.21***	0.07	0.19*	0.11	0.16	0.10
Average grade (std)	0.25***	0.07	0.20*	0.10	0.23***	0.09
# of observations	3174		1651		1572	
# of clusters	414		370		367	
# of observations off support	246		67		130	
F-statistic of first stage	208.0***		140.9***		148.9***	
F-statistic for joint significane of LATE's	4.7***		1.3		2.7**	

*Notes: This table shows the weighted LATE estimates for students of the 3rd to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.5.2: Sensitivity balanced sample - Non-cognitive skills

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.52***	0.04	0.53***	0.04	0.51***	0.04
Emotional symptoms index (std)	-0.03	0.06	-0.01	0.08	-0.01	0.09
Conduct problems index (std)	0.12*	0.06	0.17*	0.09	0.07	0.10
Hyperactivity index (std)	-0.01	0.05	-0.08	0.09	0.11	0.08
Peer relations problems index (std)	0.10	0.06	0.33***	0.10	-0.10	0.10
Asocial behaviour index (std)	-0.08	0.07	-0.08	0.11	-0.09	0.09
Total index (std)	0.05	0.05	0.11	0.09	0.04	0.09
Emotional symptoms borderline or abnormal (bin)	-0.02	0.02	0.01	0.03	-0.04	0.04
Emotional symptoms abnormal (bin)	-0.04**	0.02	-0.01	0.02	-0.05*	0.03
Conduct problems borderline or abnormal (bin)	0.02	0.03	0.05	0.05	0.001	0.04
Conduct problems abnormal (bin)	0.05**	0.02	0.06**	0.03	0.03	0.03
Hyperactivity borderline or abnormal (bin)	0.04*	0.02	0.04	0.03	0.06***	0.02
Hyperactivity abnormal (bin)	0.02	0.02	0.03	0.03	0.01	0.02
Peer relations problems borderline or abnormal (bin)	0.01	0.02	0.09**	0.04	-0.07**	0.04
Peer relations problems abnormal (bin)	0.01	0.02	0.07**	0.03	-0.04*	0.03
Asocial behaviour borderline or abnormal (bin)	-0.01	0.02	0.02	0.03	-0.05**	0.02
Asocial behaviour abnormal (bin)	-0.02*	0.01	-0.02	0.02	-0.02	0.01
Total index borderline or abnormal (bin)	0.05**	0.01	0.11***	0.03	0.002	0.03
Total index abnormal (bin)	0.03**	0.01	0.05**	0.02	0.02	0.02
# of observations	3315		1731		1651	
# of clusters	451		399		393	
# of observations off support	243		57		119	
F-statistic of first stage	222.6***		162.2***		168.0***	
F-statistic for joint significane of LATE's	2.1***		2.7***		2.1***	

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.5.3: Sensitivity balanced sample - Motor skills

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.52***	0.04	0.53***	0.04	0.51***	0.04
Push-ups (std)	-0.05	0.06	-0.02	0.09	-0.12	0.08
Side-steps (std)	0.10*	0.05	0.07	0.06	0.09	0.08
Static stand (std)	0.04	0.06	0.06	0.09	0.03	0.07
Standing long jump (std)	0.05	0.05	0.06	0.06	-0.01	0.07
Reaction time (std)	0.02	0.05	0.02	0.06	0.04	0.08
Balancing backwards (std)	0.11*	0.06	0.02	0.09	0.22**	0.08
Line tracking mistakes (std)	-0.09	0.06	-0.09	0.09	-0.02	0.07
Line tracking mistake duration (std)	-0.09	0.06	-0.07	0.09	-0.08	0.08
Line tracking duration (std)	0.09	0.08	0.09	0.10	0.09	0.11
Inserting pins duration (std)	-0.09*	0.05	-0.11*	0.07	-0.05	0.07
Stand and reach (std)	0.23***	0.07	0.14	0.09	0.32***	0.10
# of observations	3315		1731		1651	
# of clusters	451		399		393	
# of observations off support	243		57		119	
F-statistic of first stage	225.8***		160.6***		162.8***	
F-statistic for joint significane of LATE's	2.6***		1.1		2.1**	

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.5.4: Sensitivity balanced sample - Physical activity

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.52***	0.03	0.53***	0.04	0.51***	0.04
School based:						
# of days with PE	0.84***	0.05	0.82***	0.05	0.85***	0.06
# of voluntary lessons	-0.05	0.08	-0.07	0.12	-0.02	0.10
F-statistic for joint significane of LATE's	162.1***		119.7***		100.8***	
Extracurricular:						
Physical activity in club sports in minutes	-1.49	4.17	0.54	5.92	-0.22	6.25
Physical activity in leisure sports in minutes	8.70	8.35	14.65	13.32	4.56	10.67
Physical activity in out of school in minutes	7.21	9.45	15.20	15.02	4.34	11.61
Participation club sports (bin)	-0.04	0.03	-0.03	0.04	-0.02	0.05
# of days active per week	0.32***	0.12	0.40**	0.18	0.28*	0.16
Compliance with WHO guideline (bin)	0.01	0.02	0.01	0.03	0.01	0.02
Media consumption hrs/week	-0.37	1.01	-0.67	1.48	-0.25	1.22
# of club sports	-0.02	0.05	0.01	0.07	-0.02	0.07
# of leisure sports	0.23***	0.08	0.14	0.10	0.34***	0.11
# of observations	3315		1731		1651	
# of clusters	451		399		393	
# of observations off support	243		57		119	
F-statistic of first stage	221.2***		155.7***		163.9***	
F-statistic for joint significane of LATE's	2.9***		1.8*		2.0**	

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.5.5: Sensitivity balanced sample - Health parameters

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.52***	0.03	0.53***	0.04	0.51***	0.04
BMI	0.18	0.22	-0.12	0.32	0.53*	0.32
BMI (std)	0.05	0.06	-0.03	0.09	0.15*	0.09
Overweight (bin)	-0.02	0.02	-0.04	0.03	-0.02	0.03
Obese (bin)	0.01*	0.01	0.01	0.01	0.02*	0.01
Overweight or obese (bin)	-0.01	0.02	-0.03	0.03	0.004	0.03
Weight in kg	0.75	0.71	0.30	0.97	1.22	0.97
Weight in kg (std)	0.03	0.03	0.01	0.05	0.06	0.05
Subjective health 1-5	-0.06	0.04	-0.07	0.05	-0.06	0.06
Subjective health good (bin)	0.03	0.04	0.03	0.04	0.04	0.05
Subjective health very good (bin)	-0.04	0.03	-0.05	0.04	-0.05	0.05
Resting heart rate	-1.02	0.81	-0.26	1.01	-1.36	1.19
Resting heart rate (std)	-0.08	0.07	-0.02	0.08	-0.11	0.10
Height in cm	0.46	0.55	0.87	0.78	-0.12	0.69
Height (std)	0.03	0.03	0.05	0.05	-0.01	0.04
# of observations	3315		1731		1651	
# of clusters	451		399		393	
# of observations off support	243		57		119	
F-statistic of first stage	228.5***		154.0***		167.2***	
F-statistic for joint significane of LATE's	1.2		0.6		1.0	

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

H.6: Sparse model

Table H.6.1: Sensitivity sparse model - Grades

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.53***	0.03	0.55***	0.04	0.50***	0.04
German grade (std)	0.19***	0.06	0.10	0.09	0.25***	0.08
Math grade (std)	0.15**	0.06	0.14	0.09	0.13	0.09
Average grade (std)	0.19***	0.06	0.14	0.09	0.21**	0.08
# of observations	4217		2107		2108	
# of clusters	450		397		401	
# of observations off support	102		31		73	
F-statistic of first stage	254.6***		203.3***		175.9***	
F-statistic for joint significane of LATE's	3.4		0.8		3.4**	

Notes: This table shows the weighted LATE estimates for students of the 3rd to 12th grade. Inverse Probability Tilting is used to control for class level, and school type dummy, income categories, level of parents education, being foreigner, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01

Table H.6.2: Sensitivity sparse model - Non-cognitive skills

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.58***	0.03	0.60***	0.03	0.56***	0.04
Emotional symptoms index (std)	-0.04	0.05	-0.01	0.06	-0.07	0.08
Conduct problems index (std)	0.14***	0.05	0.17***	0.06	0.08	0.07
Hyperactivity index (std)	-0.04	0.04	-0.09	0.07	0.001	0.06
Peer relations problems index (std)	0.12***	0.05	0.25***	0.07	-0.01	0.07
Asocial behaviour index (std)	-0.02	0.05	-0.09	0.08	0.03	0.07
Total index (std)	0.04	0.04	0.08	0.06	-0.004	0.07
Emotional symptoms borderline or abnormal (bin)	-0.02	0.02	0.01	0.02	-0.04	0.03
Emotional symptoms abnormal (bin)	-0.03**	0.02	-0.01	0.02	-0.06**	0.02
Conduct problems borderline or abnormal (bin)	0.05**	0.02	0.06**	0.03	0.02	0.03
Conduct problems abnormal (bin)	0.02	0.02	0.04*	0.02	0.004	0.02
Hyperactivity borderline or abnormal (bin)	0.01	0.02	0.01	0.03	0.02	0.02
Hyperactivity abnormal (bin)	0.01	0.01	0.02	0.02	-0.004	0.02
Peer relations problems borderline or abnormal (bin)	0.03*	0.02	0.09***	0.03	-0.03	0.02
Peer relations problems abnormal (bin)	0.02	0.01	0.04**	0.02	-0.01	0.02
Asocial behaviour borderline or abnormal (bin)	-0.004	0.01	0.01	0.02	-0.02	0.02
Asocial behaviour abnormal (bin)	-0.01	0.01	-0.01	0.01	-0.02*	0.01
Total index borderline or abnormal (bin)	0.03**	0.02	0.06***	0.02	-0.002	0.02
Total index abnormal (bin)	0.03**	0.01	0.04***	0.02	0.02	0.02
# of observations	5326		2674		2668	
# of clusters	501		446		448	
# of observations off support	93		27		50	
F-statistic of first stage	404.0***		341.1***		259.2***	
F-statistic for joint significane of LATE's	2.8***		3.3***		1.5*	

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, and school type dummy, income categories, level of parents education, being foreigner, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.6.3: Sensitivity sparse model - Motor skills

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.57***	0.03	0.59***	0.04	0.55***	0.04
Push-ups (std)	-0.10*	0.05	-0.14*	0.07	-0.07	0.07
Side-steps (std)	0.04	0.05	0.03	0.05	0.05	0.07
Static stand (std)	0.05	0.05	0.05	0.08	0.04	0.07
Standing long jump (std)	0.004	0.04	0.01	0.05	-0.02	0.06
Reaction time (std)	0.03	0.05	0.01	0.05	0.06	0.07
Balancing backwards (std)	0.11**	0.05	0.08	0.08	0.17***	0.06
Line tracking mistakes (std)	-0.02	0.06	-0.04	0.08	0.02	0.06
Line tracking mistake duration (std)	-0.03	0.07	-0.02	0.10	-0.01	0.06
Line tracking duration (std)	0.06	0.07	0.04	0.09	0.11	0.09
Inserting pins duration (std)	-0.06	0.05	-0.05	0.06	-0.07	0.07
Stand and reach (std)	0.16***	0.06	0.09	0.07	0.27***	0.08
# of observations	4505		2299		2233	
# of clusters	489		436		434	
# of observations off support	100		19		54	
F-statistic of first stage	351.1***		269.5***		238.4***	
F-statistic for joint significane of LATE's	2.3***		1.2		2.3***	

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, and school type dummy, income categories, level of parents education, being foreigner, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.6.4: Sensitivity sparse model - Physical activity

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.59***	0.03	0.61***	0.03	0.58***	0.04
School based:						
# of days with PE	0.87***	0.04	0.85***	0.04	0.89***	0.05
# of voluntary lessons	-0.08	0.07	-0.08	0.08	-0.07	0.08
F-statistic for joint significane of LATE's	283.8***		221.7***		187.5***	
Extracurricular:						
Physical activity in club sports in minutes	-3.26	3.08	-1.43	4.28	-5.09	4.52
Physical activity in leisure sports in minutes	4.36	6.27	9.91	9.09	-0.43	7.83
Physical activity in out of school in minutes	1.10	7.08	8.47	10.31	-5.53	9.02
Participation club sports (bin)	-0.04	0.02	-0.04	0.03	-0.04	0.04
# of days active per week	0.22**	0.10	0.16	0.13	0.27**	0.13
Compliance with WHO guideline (bin)	0.01	0.02	0.01	0.03	0.01	0.02
Media consumption hrs/week	0.58	0.63	0.87	0.95	0.32	0.81
# of club sports	-0.05	0.04	-0.01	0.05	-0.10	0.06
# of leisure sports	0.13**	0.06	0.09	0.08	0.19**	0.09
# of observations	4995		2486		2534	
# of clusters	493		437		440	
# of observations off support	73		31		17	
F-statistic of first stage	390.2***		323.8***		279.1***	
F-statistic for joint significane of LATE's	2.2**		1.0		1.9**	

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, and school type dummy, income categories, level of parents education, being foreigner, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.6.5: Sensitivity sparse model - Health parameters

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.57***	0.03	0.58***	0.04	0.56***	0.01
BMI	0.06	0.17	-0.13	0.23	0.23	0.24
BMI (std)	0.02	0.05	-0.03	0.06	0.06	0.07
Overweight (bin)	-0.01	0.01	-0.01	0.02	-0.01	0.02
Obese (bin)	0.01	0.01	0.01	0.01	0.01	0.01
Overweight or obese (bin)	-0.003	0.01	-0.003	0.02	-0.01	0.02
Weight in kg	0.18	0.56	-0.05	0.70	0.14	0.71
Weight in kg (std)	0.01	0.03	-0.002	0.03	0.01	0.03
Subjective health 1-5	-0.04	0.03	-0.05	0.04	-0.03	0.05
Subjective health good (bin)	0.02	0.03	0.03	0.03	0.01	0.04
Subjective health very good (bin)	-0.03	0.03	-0.04	0.03	-0.02	0.04
Resting heart rate	-0.41	0.62	0.63	0.82	-1.38	0.92
Resting heart rate (std)	-0.03	0.05	0.05	0.07	-0.11	0.08
Height in cm	-0.14	0.47	0.15	0.62	-0.89*	0.50
Height (std)	-0.01	0.03	0.01	0.04	-0.05*	0.03
# of observations	4713		2407		2339	
# of clusters	491		437		438	
# of observations off support	107		20		54	
F-statistic of first stage	331.2***		269.5***		247.8***	
F-statistic for joint significane of LATE's	1.0		0.6		1.3	

*Notes: This table shows the weighted LATE estimates for students of the 1st to 12th grade. Inverse Probability Tilting is used to control for class level, and school type dummy, income categories, level of parents education, being foreigner, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

H.7: No adjustment for common support

Table H.7.1: Sensitivity no common support adjustment - Grades

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.54***	0.03	0.55***	0.04	0.52***	0.04
German grade (std)	0.20***	0.05	0.12	0.09	0.24***	0.08
Math grade (std)	0.14**	0.06	0.14	0.09	0.13	0.09
Average grade (std)	0.20***	0.06	0.15*	0.09	0.21***	0.08
# of observations	4319		2138		2181	
# of clusters	454		400		402	
# of observations off support	0		0		0	
F-statistic of first stage	257.9***		196.8***		199.3***	
F-statistic for joint significane of LATE's	4.3***		0.9		3.3**	

*Notes: This table shows the weighted LATE estimates for students of the 3rd to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.7.2: Sensitivity no common support adjustment - Non-cognitive skills

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.59***	0.03	0.61***	0.03	0.57***	0.03
Emotional symptoms index (std)	-0.03	0.05	-0.01	0.06	-0.07	0.08
Conduct problems index (std)	0.13***	0.05	0.17***	0.06	0.08	0.07
Hyperactivity index (std)	-0.03	0.04	-0.07	0.07	0.01	0.06
Peer relations problems index (std)	0.08*	0.05	0.23***	0.07	-0.05	0.07
Asocial behaviour index (std)	-0.04	0.05	-0.10	0.08	0.01	0.07
Total index (std)	0.04	0.04	0.08	0.06	-0.01	0.07
Emotional symptoms borderline or abnormal (bin)	-0.02	0.02	0.01	0.02	-0.04	0.03
Emotional symptoms abnormal (bin)	-0.03**	0.01	-0.01	0.02	-0.06**	0.02
Conduct problems borderline or abnormal (bin)	0.05**	0.02	0.07**	0.03	0.03	0.03
Conduct problems abnormal (bin)	0.02	0.02	0.04*	0.02	0.004	0.02
Hyperactivity borderline or abnormal (bin)	0.02	0.02	0.02	0.03	0.02	0.02
Hyperactivity abnormal (bin)	0.01	0.01	0.03	0.02	-0.002	0.01
Peer relations problems borderline or abnormal (bin)	0.02	0.02	0.08***	0.03	-0.04*	0.02
Peer relations problems abnormal (bin)	0.01	0.01	0.04**	0.02	-0.04*	0.02
Asocial behaviour borderline or abnormal (bin)	-0.01	0.01	0.01	0.02	-0.02	0.02
Asocial behaviour abnormal (bin)	-0.01	0.01	-0.01	0.01	-0.01	0.01
Total index borderline or abnormal (bin)	0.03*	0.01	0.07***	0.02	-0.01	0.02
Total index abnormal (bin)	0.02**	0.01	0.04***	0.02	0.01	0.02
# of observations	5419		2701		2718	
# of clusters	504		449		448	
# of observations off support	0		0		0	
F-statistic of first stage	390.6***		333.5***		294.2***	
F-statistic for joint significane of LATE's	2.4***		2.8***		1.9**	

*Notes: This table shows the weighted LATE estimates for students of the 3rd to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.7.3: Sensitivity no common support adjustment - Motor skills

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.57***	0.03	0.58***	0.04	0.56***	0.04
Push-ups (std)	-0.06	0.05	-0.06	0.08	-0.07	0.07
Side-steps (std)	0.10**	0.05	0.09*	0.05	0.08	0.07
Static stand (std)	0.01	0.06	0.02	0.08	0.01	0.07
Standing long jump (std)	0.03	0.04	0.03	0.05	0.02	0.05
Reaction time (std)	0.05	0.05	0.04	0.05	0.06	0.07
Balancing backwards (std)	0.14***	0.05	0.11	0.08	0.18***	0.07
Line tracking mistakes (std)	-0.04	0.05	-0.04	0.08	0.02	0.07
Line tracking mistake duration (std)	-0.04	0.06	-0.02	0.09	-0.03	0.06
Line tracking duration (std)	0.13*	0.07	0.11	0.08	0.16*	0.09
Inserting pins duration (std)	-0.09*	0.05	-0.08	0.06	-0.10	0.07
Stand and reach (std)	0.17***	0.05	0.08	0.07	0.30***	0.08
# of observations	4605		2318		2287	
# of clusters	492		439		435	
# of observations off support	0		0		0	
F-statistic of first stage	339.5***		254.9***		236.8***	
F-statistic for joint significane of LATE's	3.2***		1.3		2.4***	

*Notes: This table shows the weighted LATE estimates for students of the 3rd to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.7.4: Sensitivity no common support adjustment - Physical activity

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.59***	0.03	0.60***	0.04	0.58***	0.03
School based:						
# of days with PE	0.83***	0.03	0.82***	0.04	0.84***	0.05
# of voluntary lessons	-0.01	0.06	-0.02	0.09	0.01	0.07
F-statistic for joint significane of LATE's	292.7***		210.0***		180.4***	
Extracurricular:						
Physical activity in club sports in minutes	-0.77	3.16	1.78	4.48	-2.60	4.40
Physical activity in leisure sports in minutes	3.73	6.20	9.66	9.59	-0.75	7.81
Physical activity in out of school in minutes	2.95	6.97	11.44	10.84	-3.36	8.94
Participation club sports (bin)	-0.02	0.02	-0.02	0.03	-0.02	0.04
# of days active per week	0.18*	0.10	0.14	0.15	0.22*	0.12
Compliance with WHO guideline (bin)	-0.01	0.02	-0.02	0.03	0.002	0.02
Media consumption hrs/week	0.49	0.69	0.73	1.00	0.43	0.79
# of club sports	-0.02	0.04	0.03	0.05	-0.06	0.06
# of leisure sports	0.17***	0.06	0.12	0.08	0.23***	0.09
# of observations	5068		2517		2551	
# of clusters	496		440		440	
# of observations off support	0		0		0	
F-statistic of first stage	381.3***		306.7***		281.6***	
F-statistic for joint significane of LATE's	2.6***		1.5		2.1**	

*Notes: This table shows the weighted LATE estimates for students of the 3rd to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*

Table H.7.5: Sensitivity no common support adjustment - Health parameters

	All		Boys		Girls	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
1st stage	0.57***	0.03	0.58***	0.04	0.56***	0.04
BMI	-0.02	0.17	-0.20	0.23	0.24	0.23
BMI (std)	-0.01	0.05	-0.06	0.06	0.06	0.06
Overweight (bin)	-0.02	0.01	-0.02	0.02	-0.01	0.02
Obese (bin)	0.01	0.01	0.01	0.01	0.0	0.01
Overweight or obese (bin)	-0.01	0.02	-0.02	0.02	-0.001	0.02
Weight in kg	-0.01	0.55	-0.12	0.74	0.14	0.70
Weight in kg (std)	-0.001	0.03	-0.01	0.04	0.01	0.03
Subjective health 1-5	-0.04	0.03	-0.05	0.04	-0.02	0.05
Subjective health good (bin)	0.02	0.03	0.03	0.04	0.001	0.04
Subjective health very good (bin)	-0.02	0.03	-0.04	0.03	-0.01	0.04
Resting heart rate	-0.38	0.63	0.64	0.80	-1.39	0.95
Resting heart rate (std)	-0.03	0.05	0.05	0.07	-0.11	0.08
Height in cm	-0.02	0.45	0.49	0.62	-0.77	0.53
Height (std)	-0.001	0.03	0.03	0.04	-0.05	0.03
# of observations	4820		2427		2393	
# of clusters	494		440		438	
# of observations off support	0		0		0	
F-statistic of first stage	337.6***		248.1***		232.7***	
F-statistic for joint significane of LATE's	1.0		0.6		1.0	

*Notes: This table shows the weighted LATE estimates for students of the 3rd to 12th grade. Inverse Probability Tilting is used to control for class level, school type, and gender dummy, income categories, level of parents education, physical activity of parents, number of siblings, being foreigner, birth weight, year of birth, community size, East Germany, and educational spending per students. Inference is based on symmetric p-values of 4999 weighted bootstraps clustered at state-school type-class-wave level. * p<0.1, ** p<0.05, *** p<0.01*