Edukacja 2014, 5(130), 70–91
An interdisciplinary approach
ISSN 0239-6858

Estimating teacher effect using hierarchical linear modelling

Maciej Koniewski

Educational Value Added Section, Educational Research Institute*

This paper reports a study addressing how teacher effectiveness influences student outcomes in Polish Lower Secondary schools. Data from a Polish nationwide lower secondary school study were analysed. Data included 3883 pupils in 246 classes, in 137 schools with 202 maths teachers and 4119 pupils in 260 classes, in 143 schools with 215 language teachers. Variance of exam scores explained by teacher effect was 12% (maths) and 8% (language skills). Controlling for prior achievement, intelligence and student family background, teacher effects were 5% (maths) and 4% (language skills). Until now there has been no consensus about which teacher characteristics could explain variance in their effectiveness. The scale of "teacher authority/classroom management" explains 91% of exam scores in maths and 81% in language skills, when controlling for prior student achievement, intelligence, student family background and school location.

Keywords: sociology of education, teacher effect, school determinants of student achievement, hierarchical models, HLM.

Factors that would best explain student academic achievement have been an important topic for investigation since as early as 1902 (Rice). Since 1966 (Coleman et al.) the dominant factor favoured was student family background and social status. This claim, supported by numerous studies (Blau and Duncan, 1967; Bourdieu, 1986; Bourdieu and Passeron, 1990; Mosteller and Moynihan, 1972; Plowden, 1967), is currently disputed (Byrne et al., 2010; Hart, Petrill and Kamp Dush, 2009; Kovas, Haworth, Petrill and Plomin, 2007; McGue et al., 2007; Oliver et al., 2004). These authors suggest that after accounting for genetic factors, family influence on student achievement is much less significant than originally presumed.

Evidence based steering of educational policy relies on research of the factors that can influence effectiveness. Studies show that of the factors controlled by the school system, teacher effectiveness is the most promising in explaining academic achievement (Akiba, LeTendre and Scribner, 2007; Hanushek, 1997; Jordan, Mendro and Weerasinghe, 1997; Rivkin, Hanushek and Kain, 2005; Sanders and Rivers, 1996; Taylor et al., 2010; Wright, Horn and Sanders, 1997).

The article was written on the basis of the chapter "Teacher's characteristics" published in the report "The relevance of educational value added method for lower secondary school". The chapter was written for Educational Research Institute under the project "Research project for the development of educational value added models" co-financed by the European Union from the European Social Fund. This article was published primarily in Polish in *Edukacja, 123*(3) 2013.

* Address: Pracownia Edukacyjnej Wartości Dodanej, Instytut Badań Edukacyjnych, ul. Górczewska 8, 01-180 Warszawa, Poland. E-mail: m.koniewski@ibe.edu.pl

The article approaches the following questions: (a) What is the magnitude of teacher effect, i.e. to what extent can teachers contribute to the improvement of student academic achievement? (b) Do teachers and schools significantly vary in their capabilities to improve student academic achievement? If so, (c) Can these differences be explained by the characteristics of teachers (e.g. years of work experience, education level) and their work (e.g. teaching style)? Data used in the analysis were on maths and Polish language teachers who taught in third grades of lower secondary schools, and from their pupils, including their scores from the maths and Polish language lower secondary school exams in 2012.

Review of study results

Nye, Konstantopoulos and Hedges (2004) summarised results of 15 teacher effect estimates which were reported in five studies (Armour, 1976; Goldhaber and Brewer, 1997; Hanushek, 1971; 1992; Murnane and Phillips, 1981). Authors reported that from 7% to 21% of student academic achievement variance is explained by variation in teacher effectiveness. This result is equivalent to an effect of d = 0.32 (Cohen, 1988)[1],which means in practical terms that with an increase in the measure of teacher effectiveness by one standard deviation, mean student academic achievement increases by about one third of a standard deviation (Hattie, 2009).

Rowan, Correnti and Miller (2002), using hierarchical linear modelling, reported inter-class variance of 18–28% (maths) and 12–23% (reading) in the random intercept null model (see Bryk and Raudenbush, 1992). Taking prior student achievement, family socioeconomic status and the school's social composition into account, this amounted to between 8% and 18% (maths), and between 4% and 16% (reading) in the model.

Some researchers (e.g. Scheerens and Bosker, 1997) interpret the inter-class exam score variance in the null model as teacher effect. Such an approach is justifiable since teacher effectiveness is the strongest student achievement predictor at class level. In addition to teacher effect, at class level, peer and class size effects are also taken into account. However, they stimulate improvement of student achievement with the active participation of a teacher. Peer effect which can stimulate learning may become evident when teachers modify their teaching strategies taking pupils characteristics and potential into account (Wilkinson, 2002). In the case of small classes, slight improvement or lack of improvement in student scores can be expected unless the teacher modifies their teaching approach to use the possibilities afforded by reduced class size, adapting the curriculum and teaching methods (Robinson, 1990).

The inter-class variance is not only determined by factors at class level, but most notably, prior achievement and family background. Therefore, other researchers (e.g. Rowan, Correnti and Miller 2002) define inter-class variance as teacher effect in a model which at least controls for student family socioeconomic status and prior academic achievement.

Hierarchical modelling was used by Nye, Konstantopoulos and Hedges (2004) on data from The Tennessee Class Size Experiment known also as the STAR project (Student–Teacher Achievement Ratio). At individual student level the authors took prior achievement, gender, family socioeconomic status and ethnicity into consideration. At class level, class size, full-time teacher assistance, teacher experience and education level were taken into account. The inter-class variance in the random intercept full model was 13%

[1] Cohen's d effect is defined as the difference between the mean value of the dependent variable in the group exposed to the stimulus (e.g. effective teacher) and in the control group, divided by standard deviation, i.e. $d = \frac{x_1 - x_2}{s}$.

(maths) and 7% (reading). In the null model, it was respectively: 14% and 9%.

Even stronger evidence for the importance of teacher effect is supported by studies allowing for a genetic factor. These were commonly studies on twins. A genetic factor explains up to 65% of academic achievement (Oliver et al., 2004). Taking into consideration the genetic factor, teacher effect is estimated to be 8% (Byrne et al., 2010). As shown by Taylor et al. (2010), teaching quality moderates the genetic cognitive potential of a student. Teachers who are able to provide an optimal environment for the realisation of the natural development trajectories of their pupils, contribute the most to their educational success. A review of literature on the genetic factor in student achievement may be found in the work of Pokropek (2013).

There is no doubt that teachers, who differ in effectiveness (teaching quality), influence student scores. It is not clear, however, which specific teacher characteristics and teaching styles explain differences in teacher effectiveness as measured by student achievement. This state of affairs is reflected in the paraphrase of the popular assertion "teacher matters" made by Hattie (2009, p. 108) – "variance due to teachers makes the difference". On the other hand, as Nye, Konstantopoulos and Hedges (2004, p. 237) note "It is widely accepted that teachers differ in their effectiveness, yet the empirical evidence regarding teacher effectiveness is weak." Rivkin, Hanushek and Kain (2005, p. 419) claim that:

> There has been no consensus on the importance of specific teacher factors, leading to the common conclusion that the existing empirical evidence does not find a strong role for teachers in the determination of academic achievement and future academic and labour market success. It may be that parents and students overstate the importance of teachers, but an alternative explanation is that measurable characteristics such as teacher experience, education, and even test scores of teachers explain little of the true variation in quality.

Researchers argue about the validity of explanations of differences in teacher effectiveness by characteristics such as education, experience and salary. These controversies are very clear in the debate between Hanushek (1986; 1989; 1996; 1997) and Hedges, Laine and Greenwald (1994a; 1994b; 1996). Hanushek holds that meta-analysis of available studies indicates the lack of a strong positive relationship between teacher characteristics and student scores, at least when student family socioeconomic status is controlled for. Greenwald, Hedges and Laine believe analysis of the same data shows a positive relationship between these variables and that its strength is of practical relevance for educational policymakers. Despite lack of agreement on importance of teachers' individual characteristics and teaching styles for an explanation of variation in academic achievement, empirical evidence in favour of alternative opinions should also be considered.

Knowledge of the subject taught. In a summary of the results of 30 studies on the relationship between the level of teacher subject knowledge with student scores, 17 showed a positive relationship and 13 no relationship (Byrne, 1983). However, in studies that report lack of a relationship, such small variation in teacher subject knowledge was reported that statistically insignificant results were unavoidable. Ashton and Crocker (1987) identified 5 out of 14 studies that found a positive relationship between teacher subject knowledge and teacher effectiveness, as measured by mean pupil scores. Scores of teachers in language competence tests (verbal ability) that evaluated reading skills, grammar knowledge and vocabulary, appeared to be poorly correlated with student language competence scores (Bowles and Levin, 1968; 1992, Ehrenberg and Brewer, 1995; Ferguson and Ladd, 1996, Hanushek, 1971). Stronger relationships are present between the scores in subject exams taken by teachers and student scores in maths (Ferguson and Ladd, 1996; Rowan, Chiang and Miller, 1997). Nye, Konstantopoulos and Hedges (2004) have estimated that with an increase by one unit of teacher education measure, pupils score higher by an average of 0.037 (maths) and 0.013 (language competence) standard deviation of school achievement measures.

Wayne and Youngs (2003) reviewed studies on the relationship between characteristics of the teacher and student achievement increments. Based on the publications analysed, the authors found the significant influence of teachers' higher education on mathematical competence test scores. There is no solid evidence that this relationship exists in science, humanities or foreign languages teaching. Research by Darling-Hammond (2000) shows a strong positive impact of teachers' high qualifications of on student achievement. The author concludes that, of the variables characterising teachers, the variable "ratio of highly qualified teachers in school" is the strongest predictor of student achievement. Other variables at the school level, i.e. student–teacher ratio, class size, proportion of teachers in a school staff, in the analyses by Darling-Hammond (2000) show a positive, but very small or negligible association with student achievement. Goldhaber and Brewer (1997; 2000) and Rowan, Chiang and Miller (1997) found no relationship between student achievement and teacher education level. The authors reported the relationship between student achievement and the subject of a teacher's education. Only in the case of maths was the effect significant. Pupils, whose teachers had a master's degree in maths, achieved better scores in maths tests.

Teacher training. Evertson, Hawley and Zlotnik (1985) demonstrated the stable positive effect of formal teacher education on student achievement. As with level of subject knowledge, here there was also the problem of small variance which obscured any significant effect. Of 13 studies, 11 reported higher effectiveness of teachers with full professional qualifications over teachers without diplomas or having incomplete professional qualifications. Teaching skills were related to subject knowledge and correlated with student achievement. Many authors also found that better results were achieved by pupils of teachers who had participated in development programs for subject teaching (Wiley and Yoon, 1995; Brown, Smith and Stein, 1996).

Professional experience. A positive, but not always significant and rather non-linear relationship between professional experience and student achievement was reported (Klitgaard and Hall, 1975; Murnane and Phillips, 1981). Teachers at the start of their careers with less than three years experience are less effective than those with more. Better results are clearly achieved by pupils of teachers with more than five years experience (Rosenholtz, 1986). Teaching effectiveness decreases approaching retirement. Nye, Konstantopoulos and Hedges (2004) estimated the effect of teaching experience as 0.085 (maths) and 0.094 (language competence) standard deviation of student achievement.

Teaching atmosphere. A significant relationship between student achievement and teaching atmosphere was reported (Freiberg, 1999; Anderson, 1982). Lee and Bryk (1989) showed that a teaching atmosphere aimed at achievement was an important factor explaining differences in student scores. When teacher leadership was more focused on correcting unwanted behaviour and suggesting appropriate resolution, teaching effectiveness became greater (Brown, 2001). Mullen and Copper (1994) focused on the coherence of teacher groups as a factor positively

Table 1
Summary of meta-analysis data about teacher effect on student achievement (Hattie, 2009, p. 109)

Teacher factors with the greatest impact on student achievement	Number of meta--analyses	Number of studies	Number of achievement estimates	*D*	*SE*
Teacher training using video analysis (microteaching)	4	402	439	0.88	—
Teacher clarity	1	—	—	0.75	—
Teacher-student relationship	1	229	1 450	0.72	(0.011)
Professional development	5	537	1 884	0.62	(0.034)
Not labelling students	1	79	79	0.61	—
Quality of teaching evaluated by students	5	141	195	0.44	(0.060)
Expectations	8	674	784	0.43	(0.081)
Teacher training	3	53	286	0.11	(0.044)
Teacher subject matter knowledge	2	92	424	0.09	(0.016)
Teacher effects (total)	1	18	18	0.32	(0.020)
Total	31	2 225	5 559	0.49	(0.049)

influencing pupils achievement at school. Edmonds (1979) reported a positive relationship between student achievement and factors such as: strong leadership of the headmaster, consensus on teaching objectives and methods, high expectations of teachers towards students and using student achievement to evaluate teachers' work in the context of a secure and stable school atmosphere.

Wade (1984) in a review of studies on teacher effectiveness pointed to the significant teacher effectiveness factors such as: setting teaching objectives, length of courses, characteristics of the group being taught, school location, schedule of classes, work motivation, teaching structure and techniques. Brookover (1982), in a book summarising several studies on school effectiveness, adds the following to the list: group teaching, effective teaching strategies, effective classroom management, the use of positive reinforcement of pupils and involvement of parents in school life.

In a meta-analysis of over 100 experimental studies on teachers' work Marzano, Marzano and Pickering (2003) pointed out that the key to achieving educational success might be frequent revision, urging pupils to take notes, positive reinforcement of pupils and expression of approval when deserved, homework and individual work in class, cooperative problem solving, precise determination of objectives and providing support in their realisation, formulating hypotheses and testing them and activating the knowledge that pupils already possessed.

Hattie (2009) summarised the results of 31 meta-analyses on the impact of a teacher on student scores. Table 1 shows his results. In addition to meta-analysis of teacher effect in general, collective meta-analyses on specific characteristics of teachers and their teaching styles were presented. To report results, Hattie used Cohen's *d* measure (Cohen, 1988). In educational studies the effect of $d = 0.2$ should be interpreted as weak, $d = 0.4$ as average, $d = 0.6$ as a strong effect (Hattie, 2009, p. 9). The largest increments in student scores were attributable to teachers who participated in training on how to conduct lessons using video analysis (d = 0.88). Student achievement was improved by teachers who formulated clear and precise messages (d = 0.75). Language skills of teachers allowed them to communicate ideas in a clear and convincing way (Murnane, 1985). Those who had good relationships with their pupils were also effective (d = 0.72). Of the remaining variables which measured the teacher–student relationship, the greatest impact on student achievement followed: promoting independence, empathy, friendliness, encouraging analytical thinking, critical evaluation and creative thinking, encouraging learning and individualisation of teaching. In all these case the measure was $d > 0.4$.

While summarising the meta-analysis and review of research discussed, a positive impact on student achievement may be demonstrated by: participation by teachers in training courses, formulating clear and precise messages to students, subject matter knowledge, especially in the case of maths, formal teacher training; teaching experience; teacher group coherence; consensus on teaching objectives and methods; high expectations towards students; using student achievement to evaluate teachers' work.

Estimation of teacher effect

Data on pupils and teachers from lower secondary schools were collected as part of the study, "Research on determinants for academic achievement of pupils in primary and lower secondary schools", which was carried out by the Educational Value Added Section (EVA Section) of the Educational Research Institute from January 2010 to June 2012 (01–04.2010 – 1st stage; 09–10.2011 – 2nd stage; 05–06.2012 – 3rd stage). The study covered 5249 pupils and 3579 teachers in 292 classes in 150 schools. The sampling frame was based on information from the Education Information System and excluded: schools with a different curriculum or teaching structure, schools run by non-government entities and public schools with only one class of fewer than 10. Stratified, proportional, two-stage sampling was used. Classes were drawn from the schools. The study included all pupils from the classes selected who were at school at the time of the study. The sample was divided into 6 strata, by size of town and according to whether a school had only one or several classes.

The measurement tools used in the study were developed by the EVA Section independently or on the basis of the tools developed as part of Polish and international research projects. The results presented in this article are based on the OECD PISA survey in terms of the "household wealth" (HOMEPOS) indicator and the "teacher authority/classroom management" scale. Raven's matrices test was used as a measure of pupil intelligence.

Pupils participating in the 3rd stage of the study were assigned teachers who taught other classes for at least one year before the study. This was their last year of study in lower secondary school. Assignment of teachers to classes was carried out separately for maths and Polish language teachers. The classes which one year preceding the date of the study had had more than one maths or Polish language teacher were removed from the analysis.

Analyses were carried out on two data sets. In the first set, the scores in the maths exam were the dependent variable. The data was collected from 3883 pupils in 246 classes in 137 schools and taught by 202 maths teachers. In the second set, the scores in the Polish language exam were the dependent variable. The data was from 4119 pupils in 260 classes in 143 schools and taught by 215 Polish language teachers. Missing data imputation was not carried out and missing values were left in the database. For factor scores calculation, pair wise deletion of missing data was used. All analyses were carried out, weighting separately for the level of pupils, classes and schools. For the reporting of scale properties, analyses were performed on unweighted data.

The results should be reported while accounting for possible threats to validity from both study scheme and method of data analysis. The results may be subject to bias because: (a) parents may choose where they live and therefore select the network of schools available for their children, (b) parents may choose a specific school for their children from outside the normal choice which they are assigned by place of residence, (c) the headmaster or administration may assign pupils and teachers to specific classes on the basis of explicit or implicit rules or requests by parents. Results reported here may therefore be biased by auto-selection (parents choosing schools for their pupils) and selection (explicit and implicit mechanisms of enrolment) of pupils in schools and classes (Dolata, 2008; Jordan, Mendro and Weerasinghe, 1997; Sanders and Rivers, 1996).

In order to reduce the potential bias, arising from non-random assignment of pupils to teachers of varying effectiveness levels, the characteristics of pupils and their family background were taken into account in the analyses. The prior achievements of pupils in classes were also taken into consideration, controlling for possible segregation of pupils in classes according to prior achievements. The mean socioeconomic status of pupils in classes was also taken into account, controlling for social composition of a class (i.e. possible segregation of pupils in classes for reasons of socioeconomic status).

Analyses were performed using hierarchical linear modelling. In such hierarchical models exogenous variables are present at different levels, most often at two levels (e.g. pupils, school) or three levels (e.g. pupils, classes, schools). In the extensive literature on multilevel models, special attention should be paid to Raudenbush and Bryk (2002). Extensive information may be also found in the HLM7 software documentation with which the analyses were performed, available at ssicentral.com. Domański and Pokropek (2011) and Domański et al. (2012) provide detailed knowledge about hierarchical models.

Results for maths teaching

Table 2 presents the coefficients and standard errors for models using independent variables at the level of pupils, classes and schools. The *p*-values are not reported. They may be calculated by dividing the coefficient of a given variable by its standard error, and then reading the critical value from the normal distribution or using available calculators.

On the basis of an intercept in the null model we may predict the maths test score of a student. The average student scaled test score was 0.06 standard deviation. Raw results were scaled with the 3 parametric logistic model (3PLM) (Jakubowski and Pokropek, 2009). The results are reported on a scale with a mean 0 and standard deviation 1. The mean in the sample differs from the intercept in the model due to missing data.

Intra class correlation (ICC) may be interpreted as a percentage of the exam scores explained variance. In the null model that only accounts for the fact that the pupils are grouped in classes and schools, the ICC was 0.121. This means that grouping of pupils in classes explained 12.1% of the maths exam score variance. Whereas grouping of pupils in schools explained 9.1% of the maths exam score variance.

Characteristics of the pupils and their families were added to the null model, i.e. prior academic achievement (exam results after primary school), gender, permanent dyslexia (statement of a student's dyslexia at a test after sixth grade and after ninth grade), intelligence (measured with Raven's matrices test), parents' education level (the duration in years of the higher educated parent's education), number of people in the household, household wealth. The resulting

EVA model[2] of schools (1MAT)[3] explained 58% of the maths exam scores variance at student level.

Residuals of the EVA (1MAT) model are a measure of school effectiveness for maths teaching. School effectiveness is mainly the result of teachers' work, school organisational culture, atmosphere at work and infrastructure etc. To obtain the measure of teacher effectiveness, effects other than teaching need to be "extracted". For this purpose, school-level variables, as well as variables at the level of the municipality where a school is located, were added to the EVA (1MAT) model. Over 40 variables potentially explaining student academic achievement were investigated, including: the number of teachers, their age distribution, experience, average class size, number of inhabitants, economic structure of municipality, municipality expenditure in general and on education per capita, number of libraries per capita, size of library per capita, unemployment rate and percentage of people on social assistance. Unfortunately, the author did not have comprehensive data at school level about school finances or infrastructure. The problem with the availability of good quality data on operation of schools was owing to the lack of an effective system for collecting and monitoring a wide range of school characteristics. Good quality data, however, were available about municipalities.

Of more than 40 variables tested, the location, measured on a four-level ordinal scale (village, town up to 20 000 inhabitants, city of more than 20 000 inhabitants, city from 20 000 to 100 000 inhabitants, city of more than 100 000 inhabitants) was an important variable explaining maths exam scores at school level. The correlation between a lower secondary school location with an average maths exam score in a municipality was 0.393. With student-level variable control, i.e. prior academic achievement, intelligence, parents' education level, the location effect on student achievement was negative. The larger the city, the lower the maths exam scores, while controlling for prior achievement, intelligence and status variables.

Dynamics of unemployment were another important maths score predictor at school level, measured as the year to year mean change in the share of the unemployed in the number of people at working age where the school was located. The more unemployment rate rose, the lower the scores achieved by pupils.

After additional inclusion in the EVA (1MAT) school model of location, unemployment dynamics and mean student test scores in a class, residuals obtained from the 2MAT model could be used as an accurate teacher effectiveness indicator. A classes average test score was included in order to control for the potential streaming of pupils on the basis of prior academic achievement. The average level of status variables in a class however, proved unhelpful and they were not employed in the model.

It should be noted that in the EVA (2MAT) model only the variables included (observed) in the study were separated. Complete separation of factors other than teacher-related variables could only be achieved in a controlled experiment. Therefore, it could be argued that the resulting indicator was more

[2] The proposed EVA model of schools differs from the EVA models used in Poland. The EVA school indicators in the Polish version are computed from two-level models which account for pupils' external exam scores at the end of the previous stage of education (after sixth grade), gender and dyslexia. In the case of EVA calculators available for teachers and school headmasters, these are one-level models. Extensive materials on the EVA indicators used in Poland are available at ewd.edu.pl and as conference materials of the Polish Association of Educational Diagnostics (including Pokropek, 2009; Pokropek and Żółtak, 2012). The Polish EVA indicators should be interpreted as indicators of school efficiency. By adding the class level, i.e. taking into account grouping of pupils in classes, and school-level variables, allows discussion of teacher efficiency.

[3] The symbols in parentheses in the text correspond to the symbols in the headings of Tables 2 and 4, which are intended to facilitate identification of the described models.

Table 2

Determinants of mathematical achievement. Results of the three-level linear model, random intercepts models, robust estimation of standard errors

Dependent variable: lower secondary school maths exam score (standardisation *Z*) (ZIRTMAT)[a]

	Null	(*SE*)	(1MAT)	(*SE*)	(2MAT)	(*SE*)	(3MAT)	(*SE*)	(4MAT)	(*SE*)
Fixed effects estimation										
STUDENT LEVEL										
Test (std. *Z*) (ZSP3PL)			0.55*	0.018	0.55*	0.019	0.55*	0.019	0.55*	0.019
Gender[b] (EVA_PLEC)			-0.12*	0.019	-0.12*	0.019	-0.12*	0.019	-0.13*	0.019
Dyslexia at a primary school exit test and the lower secondary school exit exam[c] (DYS)			-0.18*	0.043	-0.18*	0.043	-0.18*	0.043	-0.18*	0.043
Intelligence (Raven's matrices test) (std. *Z*) (RAVEN)			0.30*	0.017	0.30*	0.017	0.30*	0.017	0.29*	0.017
Higher of the number of years of parent's education (HEDULL)			0.02*	0.005	0.02*	0.005	0.02*	0.005	0.02*	0.005
Household wealth (std. *Z*) (HOMEPOS)			0.04*	0.016	0.03*	0.016	0.04*	0.016	0.03*	0.016
Number of people in a household (WIELRODZ)			0.02*	0.007	0.02*	0.007	0.02*	0.007	0.02*	0.007
Intercept	0.06	0.039	-0.11	0.077	-0.12	0.080	-0.12	0.080	-0.14	0.080
CLASS LEVEL										
Scale "active teaching style" (std. *Z*) (ACTIVM5)							0.10*	0.032		
Scale "teacher authority/classroom management" (std. *Z*) (DYSM)									0.13*	0.025
Average student scores at a test in class (std. *Z*) (ZSPRAWD2)					0.14*	0.041	0.17*	0.040	0.13*	0.038
SCHOOL LEVEL										
Lower secondary school location (LOK)					-0.05*	0.018	-0.05*	0.017	-0.04*	0.017
Average unemployment dynamics YoY in % (2006–2010) (YOYBEZ)					-0.86*	0.179	-0.81*	0.166	-0.78*	0.171

Random effects estimation									
School effects variance	0.09*	0.03*		0.02*		0.02*		0.02*	
Class effects variance (% of explained variance)	0.12*	0.02*		0.02*	(86%)	0.02*	(86%)	0.01*	(91%)
Student-level variance (% explained variance)	0.77*	0.32*	(58%)	0.32*		0.32*		0.32*	
Intra-class correlation coefficient[d]									
School level	0.091	0.074		0.061		0.054		0.057	
Class level	0.121	0.053		0.045		0.045		0.031	
Total school level and class level	0.211	0.127		0.105		0.099		0.088	
Summary									
Deviance	10 401.07	6 920.89		6 888.52		6 880.08		6 858.71	
Number of parameters estimated	4	11		14		15		15	
Comparison of the model's fit to the null model	–	33.46%		33.77%		33.85%		34.06%	
Improvement of fit by adding predictors	–	33.46%		0.47%		0.12%		0.43%	

Number of pupils: $n = 3883$, number of classes: $n = 246$, number of schools: $n = 137$ (analysis on weighted data).

* Effect is significant at 0.05 level.

[a] In parentheses are the names of variables, to facilitate reading of the equations in the Appendix 2.

[b] Reference group: boys.

[c] Reference group: pupils without dyslexia.

[d] In the case of models with predictors for the intra-class correlation coefficients, the same interpretation as for the null model was used. These values constitute % of the student scores of the variance to be explained at class level, after taking into account student-level predictors.

Table 3
Correlations of residual variable of the (2MAT) model with the exam scores and selected potential predictors of teacher effectiveness in maths teaching.

Selected predictors of teacher effectiveness	Teacher effectiveness	Teacher effectiveness after separation of class level factors
Average maths exam score (std. *Z*)	0.375*	0.359*
Teacher's age	0.100	0.094
Professional experience in education	0.071	0.080
Professional advancement	0.067	0.157*
Active teaching style (scale) (std. *Z*)	0.128*	0.144*
Teacher authority/classroom management (scale) (std. *Z*)	0.323*	0.317*
Number of pupils in a class	0.031	0.014

Number of teachers: n = 246 (analysis on unweighted data).
* Correlation is significant at 0.05 level.

due to the characteristics of a class than to teacher effectiveness. This may be tested.

If the EVA (2MAT) model for the maths exam were to be expanded by the EVA indicator for Polish language teachers (obtained from the analogical 2POL model), and the EVA (2POL) model for the Polish language extended by the EVA maths teacher indicator, then in such models the possible impact on student achievement of factors other than teacher-related factors at class level would be controlled. The EVA coefficient for Polish language teachers in the EVA model for the maths exam (2MAT) was 0.34 ($SE = 0.14$) and was statistically significant ($p < 0.05$). Whereas the EVA coefficient for maths teachers in the EVA model for Polish language teachers (2POL) was 0.26 ($SE = 0.17$) and was statistically insignificant ($p < 0.05$). In either case, these variables should not be statistically significant. It may, therefore, be concluded that the proposed EVA model for maths teachers is inferior to the EVA model for Polish language teachers as the EVA indicator for maths teachers is in fact a measure for both teacher effectiveness and other class characteristics, such as the atmosphere in a classroom, peer effect, etc. Table 3 shows the correlation between residual variables of the 2MAT model and the model extended by the EVA indicator of maths teachers and selected potential teacher effectiveness predictors.

As expected, after separation of class factors from teacher effect, higher correlations were observed for teacher characteristics and their teaching style (experience, degree of professional advancement, the "active teaching style" scale), and lower correlations for variables related in total (number of pupils) or in part (exam scores, scale "authority.../ /management...") with class characteristics.

An attempt was made to determine which teacher and teacher work characteristic explained inter-class differences in teacher effectiveness. 26 independent variables were tested, including 14 teacher characteristics related to experience, level of advancement and professional development. Included among variables tested, there were 7 scales constructed on the basis of questions from teacher questionnaires and 5 scales constructed on the basis of questions from student questionnaires. The scales related to teaching styles. The results are presented in Table 2 (3MAT and 4MAT models).

Owing to the very small variances, gender (90% of women in the sample) and teacher education level (98% of teachers in the sample had at least a master's degree) were insignificant as maths score predictors, in addition to teacher's age and professional experience, professional advancement, participation in training programmes and most of the teaching styles scales.

The "active teaching style" allowed teacher effectiveness to be described in four dimensions: "individualised learning", "use of multimedia", "revision of material", "stimulating group work". The technical documentation of the scale is presented in the Appendix 1. Teachers, who approached pupils individually, used multimedia, revised study material and organised group work, improved pupil achievement by a mean of 0.10 standard deviation from the maths score. A stronger predictor which eliminated the significance of the "active teaching style" scale is the "teacher authority/classroom management" scale. The technical documentation for the scale is presented in the Appendix 1. Pupils scored an average of 0.13 standard deviation higher in the maths exam with an increase of one *SD*. on this scale. The "teacher authority/ /classroom management" scale explained 91% of teacher effect variance in maths with control for prior academic achievement, gender, dyslexia, intelligence, student's family status variables, school location and dynamics of unemployment. Although, after considering predictors at student level, there is little remaining variance to be explained at class level. The percentage value is the so-called pseudo R^2 (Raudenbush and Bryk, 2002)[4].

Based on *Deviance* statistics, the quality of the model fit may be assessed. The lower the *Deviance* statistic, the better the fit. By adding "active teaching style" scale and "average exam score in a class" to the variables at individual level the fit is improved by 0.59% (3MAT model). Using the "teacher authority/ /classroom management" scale rather than the "active teaching style" scale improved the fit by 0.90% (4MAT model).

Results for Polish language teaching

Table 4 presents the coefficients and standard errors for models with the Polish language exam scores as dependent variables. Each student scored an average of 0.07 standard deviation of the scaled test score. Grouping of pupils in classes explained 8.2%, and grouping of pupils in schools explained 7.8% of the Polish language exam scores variance.

Characteristics of the pupils and their families were added to the null model, i.e. prior academic achievement, gender, intelligence, parents' education level, number in the household and household wealth. Permanent dyslexia is a non-significant Polish language exam score predictor at pupil level. The resulting EVA model of schools (1POL) explained 55% of the Polish language score variance at student level.

A significant Polish language exam score predictor at school level is location of the lower secondary school. As in the case of maths, with control of variables at student level, i.e. prior academic achievement, intelligence and parents' education level, the effect of location on student scores was negative. The larger the city, the lower was the Polish language score with control for prior achievement, intelligence and status.

The dynamics of unemployment significantly predicted maths exam scores but were not significant for the Polish language exam. Average test scores as an aggregated variable at class level were insignificant. After further taking into consideration of municipality location in the EWD (1POL) model, the residuals obtained from the model (2POL) may be regarded as a rough indicator for teacher effectiveness.

[4] Pseudo R^2 measures the degree of reduction of unexplained variance at a given level of analysis (pupils, classes, schools) in the null model compared to the model with explanatory variables. However, in interpretation of this measure, caution is advised (Kreft and de Leeuw, 1998).

Table 4
Determinants of language achievement. Results of the three-level linear model, random intercepts models, robust estimation of standard errors

Dependent variable: Polish language lower secondary school exam score (standardisation *Z*) (ZIRTPOL)										
	Null	(*SE*)	(1POL)	(*SE*)	(2POL)	(*SE*)	(3POL)	(*SE*)	(4POL)	(*SE*)
Fixed effects estimation										
STUDENT LEVEL										
Test (std. *Z*) (ZSP3PL)			0.54*	0.02	0.54*	0.02	0.54*	0.02	0.54*	0.02
Gender[(a)] (EVA_PLEC)			0.29*	0.02	0.29*	0.02	0.28*	0.02	0.28*	0.02
Intelligence (Raven's matrices test) (std. *Z*) (RAVEN)			0.17*	0.02	0.17*	0.02	0.17*	0.02	0.16*	0.02
Higher of the number of years of parents' education (HEDULL)			0.03*	0.01	0.03*	0.01	0.03*	0.01	0.03*	0.01
Household wealth (std. *Z*) (HOMEPOS)			0.08*	0.02	0.08*	0.02	0.08*	0.02	0.08*	0.02
Number of people in a household (WIELRODZ)			-0.02*	0.01	-0.03*	0.01	-0.03*	0.01	-0.03*	0.01
Intercept	0.07*	0.03	-0.31*	0.08	-0.23*	0.08	-0.23*	0.08	-0.23*	0.08
CLASS LEVEL										
Scale "helpful teacher" (std. *Z*) (POMOCNP)							0.07*	0.03		
Scale "teacher authority/classroom management" (std. *Z*) (DYSP)									0.11*	0.03
SCHOOL LEVEL										
Lower secondary school location (LOK)					-0.04*	0.02	-0.04*	0.02	-0.04*	0.02
Random effects estimation										
School effects variance	0.07*		0.02*		0.02*		0.02*		0.02*	
Class-level variance (% explained variance)	0.08*		0.02*		0.02*	(79%)	0.02*	(80%)	0.01*	(81%)
Student-level variance (% explained variance)	0.78*		0.35*	(55%)	0.35*		0.35*		0.35*	

Intra-class correlation coefficient					
School level	0.078	0.057	0.051	0.051	0.043
Class level	0.082	0.042	0.042	0.039	0.039
Total school level and class level	0.160	0.100	0.093	0.090	0.082
Summary					
Deviance	11 033.91	7 615.64	7 605.77	7 599.31	7 586.19
Number of parameters estimated	4	10	11	12	12
Comparison of the model's fit to the null model	–	30.98%	31.07%	31.13%	31.25%
Correction of fit by adding predictors	–	30.98%	0.13%	0.08%	0.26%

Number of pupils: n = 4119, number of classes: n = 260, number of schools: n = 143.
[*] Effect is significant at 0.05 level.
[(a)] Reference group: boys.

Table 5

Correlations of residual variable of the (2POL) model with the exam results and selected potential predictors of teacher effectiveness in Polish language teaching.

Selected teacher effectiveness predictors	Teacher effectiveness	Teacher effectiveness after separation of class level factors
Average Polish language exam score (std. *Z*)	0.524*	0.485*
Teacher's age	0.050	-0.040
Professional experience in education	-0.078	-0.084
Professional advancement	0.086	0.029
Helpful teacher (scale) (std. *Z*)	0.176*	0.192*
Teacher authority/classroom management (scale) (std. *Z*)	0.231*	0.246*
Number of pupils in a class	0.024	0.016

Number of teachers: $n = 246$ (analysis on unweighted data).
* Correlation is significant at 0.05 level.

Table 5 shows the correlation between residual variables of the 2POL model and the model extended by the EVA indicator for Polish language teachers and selected potential teacher effectiveness predictors.

A significant and high (0.485–0.524) correlation for teacher effectiveness with pupils' Polish language exam scores confirmed the hypothesis that an effective teacher contributed significantly to the educational success of their pupils. The best teacher effectiveness predictor for Polish language teaching was – as in maths teaching – the "teacher authority/classroom management" scale (0.231–0.246).

As with maths teaching, with Polish language teaching, an attempt was made to determine which teacher and work characteristics might explain inter-class differences in teacher effectiveness. The results are presented in Table 4 (models 3POL and 4POL).

Due to very minor variance, gender (97.5% of women in the sample) and teacher education level (98% of teachers with at least higher education and master's degree in the sample) proved to be insignificant Polish language exam score predictors. Teacher age and teaching experience, professional advancement and participation in training programs also proved not to be significant Polish language score predictors.

The "helpful teacher" scale significantly describes teacher effectiveness. With an increase of one standard deviation on the "helpful teacher" scale, lower secondary school pupils scored on average 0.07 standard deviations higher in the Polish language exam. A stronger predictor which overwhelms the "helpful teacher" scale's significance is the "teacher authority/classroom management" scale. With an increase of one standard deviation on the "teacher authority/classroom management" scale, lower secondary school pupils scored on average 0.11 standard deviations higher in the Polish language exam. The "teacher authority/classroom management" scale explains 81% of teacher effect variance, controlling for prior academic achievement, gender, dyslexia, intelligence, family status, school location and the dynamics of unemployment in the municipality.

Conclusions

The student score variance estimates explained by teacher effectiveness found in the study were consistent with other results in the literature. Exam score variances attributable to teachers were 12% (maths) and 8% (language competence). Rowan, Correnti and Miller (2002) reported 18–28% (maths) and 12–23% (language competence) student score variance explained by teacher effect. Nye, Konstantopoulos and Hedges (2004) found: 13–14% and 7–9%.

Teacher effect in the model explains 5% (maths) and 4% (language competence) of exam score variance, accounting for prior achievement, gender, dyslexia, intelligence and family status. Rowan, Correnti and Miller (2002) report respectively 8–18% and 4–16%. The reported effect is of great practical relevance compared with the size of the effects of other factors at class and school level, such as class size (Koniewski, 2012).

An effective teacher has a greater impact on the learning process than any other single factor controlled by the school system (Wright, Horn and Sanders, 1997; Rivkin, Hanushek and Kain, 2005). The teacher should therefore be the central focus for educational policy. If we want to improve the education system, we should start with teaching skills (Hanushek, 1997; Akiba, LeTendre and Scribner, 2007).

The "teacher authority/classroom management" scale has proven to be a stable and strong teacher effectiveness predictor both in maths and Polish language teaching. Teachers who are able to maintain discipline in the classroom, interest pupils in a subject and capture their attention, improve student scores by an average of 0.13 standard deviation in maths and 0.11 in the Polish language. The "teacher authority/classroom management" scale explains 91% of teacher effect variance in maths teaching and 81% in Polish language teaching, controlling for prior academic achievement, gender, intelligence, family status variables and location.

The results obtained confirm that teacher effect estimation based on Polish data is consistent with findings from data collected in the United States. Also noteworthy is the degree of consistency with the results reported by Nye, Konstantopoulos and Hedges (2004), who analysed high-quality experimental data.

Literature

Akiba, M., LeTendre, G. K. and Scribner, J. P. (2007). Teacher quality, opportunity gap, and national achievement in 46 countries. *Educational Researcher*, *36*(7), 369–387.

Anderson, C. S. (1982). The search for school climate – a review of the research. *Review of Educational Research*, *52*(3), 368–420.

Armour, D. T. (1976). *Analysis of the school preferred reading program in selected Los Angeles minority schools*. Santa Monica, CA: Rand Corporation.

Ashton, P. and Crocker, L. (1987). Systematic study of planned variations – the essential focus of teacher – education reform. *Journal of Teacher Education*, *38*(3), 2–8.

Blau, P. M. and Duncan, O. D. (1967). *The American occupational structure*. New York: John Wiley and Sons.

Bourdieu, P. (1986). The forms of social capital. In J. Richardson (ed.), *Handbook of theory and research for the sociology of education* (pp. 241–258). New York, NY: Greenwood Press.

Bourdieu, P. and Passeron, J. C. (1990). *Reprodukcja: elementy teorii systemu nauczania* [*Reproduction: elements of the theory of the educational system*]. Warszawa: PWN.

Bowles, S. S. and Levin, H. M. (1968). More on multicollinearity and the effectiveness of schools. *The Journal of Human Resources*, *3*(3), 393–400.

Brookover, W. B. (1982). *Creating effective schools: an in-service program for enhancing school learning climate and achievement*. Holmes Beach, FL: Learning Publications.

Brown, C., Smith, M. and Stein. M. (1996). Linking teacher support to enhanced classroom instruction. Paper presented at the conference of tThe American Educational Research Association, New York, NY.

Brown, D. F. (2001). The value of advisory sessions for urban young adolescents. *Middle School Journal*, *32*(4), 14–22.

Bryk, A. S., and Raudenbush S. (1992). *Hierarchical linear models in social and behavioral research: applications and data analysis methods*. Newbury Park, CA: Sage Publications.

Byrne, B., Coventry, W. L., Olson, R. K., Wadsworth, S. J., Samuelsson, S., Petrill, S. A., Willcutt, E. G. and Corley, R. (2010). „Teacher effects" in early literacy development: evidence from a study of twins. *Journal of Educational Psychology*, *102*(1), 32–42.

Byrne, C. (1983). Teacher knowledge and teacher effectiveness: a literature review, theoretical analysis and discussion of research strategy. Paper presented at the conference Northwestern Educational Research Association, Ellenville, NY.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: L. Erlbaum.

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

Darling-Hammond, L. (2000). Teacher quality and student achievement: a review of state policy evidence. *Education Policy Analysis Archives*, *1*(8), 1–44.

Dolata, R. (2008). *Szkoła – segregacje – nierówności* [*School – segregation – inequality*]. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Domański, H. and Pokropek, A. (2011). *Podziały terytorialne, globalizacja a nierówności społeczne. Wprowadzenie do modeli wielopoziomowych* [*The territorial divisions, globalization and social inequality. An introduction to multilevel models*]. Warszawa: Wydawnictwo IFiS PAN.

Domański, H., Karpiński, Z., Pokropek, A., Przybysz, D., Sawiński, Z., Słomczyński, K. M. and Trzciński, R. (2012). *Metodologia badań nad stratyfikacją społeczną* [*The methodology of social stratification research*]. Warszawa: Scholar.

Edmonds, R. (1979). Effective schools for urban poor. *Educational Leadership*, *37*(1), 15–24.

Ehrenberg, R. G. and Brewer, D. J. (1995). Did teachers' verbal ability and race matter in the 1960's? Coleman revisted. *Economics of Education Review*, *14*(1), 1–21.

Evertson, C. M., Hawley, W. D. and Zlotnik, M. (1985). Making A difference in educational quality through teacher – education. *Journal of Teacher Education*, *36*(3), 2–12.

Ferguson, R. F. and Ladd, H. L. (1996). How and why money matters: an analysis of Alabama schools. In H. F. Ladd (ed.), *Holdine schools accountable*. Washington, D.C.: The Brookings Institution.

Forero, C. G., Maydeu-Olivares, A. and Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: a Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling A Multidisciplinary Journal*, *16*(4), 625–641.

Freiberg, H. J. (ed.). (1999). School climate: measuring, improving and sustaining healthy learning environments. Philadelphia, PA: Falmer Press.

Goldhaber, D. D. and Brewer, D. J. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources*, *32*(3), 505–523.

Goldhaber, D. D. and Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis*, *22*(2), 129–145.

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement – estimation using micro data. *American Economic Review*, *61*(2), 280–288.

Hanushek, E. A. (1986). The economics of schooling production and efficiency in public schools. *Journal of Economic Literature*, *24*(3), 1141–1177.

Hanushek, E. A. (1989). The impact of differential expenditures on school performance. *Educational Researcher*, *18*(4), 45–62.

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy*, *100*(1), 84–117.

Hanushek, E. A. (1996). A more complete picture of school resource policies. *Review of Educational Research*, *66*(3), 397–409.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: an update. *Educational Evaluation and Policy Analysis*, *19*(2), 141–164.

Hart, S. A., Petrill, S. A. and Thompson, L. A. (2009). A factorial analysis of timed and untimed measures of maths and reading abilities in school aged twins. *Learning and Individual Differences*, *20*(2), 63–69.

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

Hedges, L. V., Laine, R. D. and Greenwald, R. (1994a). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher*, *23*(3), 5–14.

Hedges, L. V., Laine, R. D. and Greenwald, R. (1994b). Money does matter somewhere: a response to Hanushek. *Educational Researcher*, *23*(4), 9–10.

Hedges, L. V., Laine, R. D. and Greenwald, R. (1996). The effect of school resources on student achievement. *Review of Educational Research*, *66*(3), 361–396.

Jakubowski, M., Pokropek, A. (2009). *Badając egzaminy. Podejście ilościowe w badaniach edukacyjnych* [*Analysing the exams. The quantitative approach in educational research*]. Warszawa: Centralna Komisja Egzaminacyjna.

Jordan, H. R., Mendro, R. and Weerasinghe, D. (1997). Teacher effects on longitudinal student achievement: a preliminary report on research on teacher effectiveness. Paper presented in The National Evaluation Institute, Indianapolis, IN.

Klitgaard, R. E. and Hall, G. R. (1975). Are there unusually effective schools? *Journal of Human Resources*, *10*(1), 90–106.

Koniewski, M. (2012). Szacowanie wpływu liczebności klasy na osiągnięcia edukacyjne uczniów z wykorzystaniem eksperymentu ex post facto [Estimating the effect of class size on academic achievement by ex post facto experiment]. *Edukacja*, *117*(1), 23–43.

Kovas, Y., Haworth, C. M. A., Petrill, S. A. and Plomin, R. (2007). Mathematical ability of 10-year-old boys and girls: genetic and environmental etiology of typical and low performance. *Journal of Learning Disabilities*, *40*(6), 554–568.

Kreft, I. and Leeuw, J. de (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage Publications.

Lee, V. E. and Bryk, A. S. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education*, *62*(3), 172–192.

Marzano, R. J., Marzano, J. S. and Pickering, D. J. (2003). *Classroom management that works*. Alexandria, VA: ASCD.

McDonald, R. P., Ho, M-H. (2002). Principles and Practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64–82.

McGue, M., Keyes, M., Sharma, A., Elkins, I., Legrand, L., Johnson, W. and Iacono, W. G. (2007). The environments of adopted and non-adopted youth: evidence on range restriction from the Sibling Interaction and Behavior Study (SIBS). *Behavior Genetics*, *37*(3), 449–462.

Mosteller, F. and Moynihan, D. P. (1972). On equality of educational opportunity. New York: Random House.

Mullen, B. and Copper, C. (1994). The Relation between group cohesiveness and performance – an integration. *Psychological Bulletin*, *115*(2), 210–227.

Murnane, R. J. (1985). Do effective teachers have common characteristics: interpreting the quantitative research evidence. Paper presented at The National Research Council Conference on Teacher Quality in Science and Maths, Washington, D.C.

Murnane, R. J. and Phillips, B. R. (1981). What Do effective teachers of inner-city children have in common. *Social Science Research*, *10*(1), 83–100.

Nestler, S. (2013). A Monte Carlo study comparing PIV, ULS and DWLS in the estimation of dichotomous confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 127–143.

Nye, B., Konstantopoulos, S. and Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257.

Oliver, B., Harlaar, N., Hayiou, T., Hayiou, M. E., Kovas, Y., Walker, S. O., Petrill, S. A., Spinath, F. M., Dale, P. S. and Plomin, R. (2004). A twin study of teacher – reported maths performance and low performance in 7-year-olds. *Journal of Educational Psychology*, *96*(3), 504–517.

Plowden, B. (1967). Children and their primary schools. London: HMSO.

Pokropek, A. (2009). Metody statystyczne wykorzystywane w szacowaniu trzyletnich wskaźników egzaminacyjnych [The statistical methods used in estimating the three-year indicators of examination]. In B. Niemierko and M. K. Szmigel (eds.), *Badania międzynarodowe and wzory zagraniczne w diagnostyce edukacyjnej* [*International studies and foreign models in educational diagnosis*] (pp. 137–153). Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Pokropek, A. (2013). *Efekt rówieśników w nauczaniu szkolnym* [*The effect of peers in school education*]. Warszawa: Instytut Badań Edukacyjnych.

Pokropek, A. and Żółtak, T. (2012). Nowe modele jednorocznej EWD [New models of one-year educational value added]. In B. Niemierko and M. K. Szmigel (eds.), *Regionalne i lokalne diagnozy edukacyjne* [*Regional and local educational diagnoses*] (pp.178–187). Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks: Sage Publications.

Rice, J. M. (1902), Educational research: a test in arithmetic. *The Forum*, October, 281 –297.

Rivkin, S. G., Hanushek, E. A. and Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica*, *73*(2), 417–458.

Robinson, G. E. (1990). Synthesis of research on the effects of class size. *Educational Leadership*, *47*(7), 80–90.

Rosenholtz, S. J. (1986). Organizational conditions of teacher learning. *Teaching and Teacher Education*, *2*(2), 91–104.

Rowan, B., Chiang, F. S. and Miller, R. J. (1997). Using research on employees' performance to study the effects of teachers on students' achievement. *Sociology of Education*, *70*(4), 256–284.

Rowan, B., Correnti, R. and Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: insights from the prospects study of elementary schools. *Teachers College Record*, *104*(8), 1525–1567.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Psychometric Monograph No. 17. Richmond, VA: Psychometric Society.

Sanders, W. L. and Rivers J. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Research progress report. Knoxville, TN: University of Tennessee Value Added Research and Assessment Center.

Scheerens, J. and Bosker, R. (1997). *The foundations of educational effectiveness*. New York: Pergamon.

Taylor, J., Roehrig, A. D., Hensler, B. S., Connor, C. M. and Schatschneider, C. (2010). Teacher quality moderates the genetic effects on early reading. *Science*, 328(5977), 512–514.

Wade, R. K. (1984). What makes a difference in inservice teacher education? A meta-analysis of research. *Educational Leadership*, *42*(4), 48–54.

Wayne, A. J. and Youngs, P. (2003). Teacher characteristics and student achievement gains: a review. *Review of Educational Research*, *73*(1), 89–122.

Wiley, D. E. and Yoon, B. (1995). Teacher reports on opportunity to learn-analyses of the 1993 California Learning Assessment System (CLAS). *Educational Evaluation and Policy Analysis*, *17*(3), 355–370.

Wilkinson, I. A. G. (2002). Introduction: peer influences on learning: where are they? *International Journal of Educational Research*, *37*(5), 395–401.

Woods, C. M. (2002). Factor analysis of scales composed of binary items: illustration with the Maudsley Obsessional Compulsive Inventory. *Journal of Psychopathology and Behavioral Assessment*, *24*(4), 215–223.

Wright, P. S., Horn. S. P. and Sanders. W. L. (1997). Teacher and classroom context effects on student achievement: implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *11*(1), 57–67.

Acknowledgments

The author wishes to thank the reviewers of EDUKACJA for their constructive comments and the participants of seminar lead by Prof. Jarosław Górniak at the Institute of Sociology of Jagiellonian University in Cracow.

Appendix 1 – scales

To create the "teacher authority/classroom management", "active teaching style" and "helpful teacher" scales, confirmatory factor analysis (CFA) was used. The analyses were performed using the Mplus software package.

Mplus default Delta parameterisation was used. Weighted least squares (WLS) estimation was used. It was decided to use the WLS method, as the method recommended in the literature for large samples ($N > 200$) and high factor loadings ($\lambda > 0.7$) (Nestler, 2013). The WLS method also allows the assessment of models fit by using χ^2 statistics in the traditional way. In the case of a small number of indicators, such as the scales presented in the article, the WLS method is a recommended method over another one, a method commonly used for categorical variables, i.e. DWLS (diagonally weighted least squares) (Woods, 2002). The DWLS tends to overestimate parameters (Forero, Maydeu-Olivares and Gallardo-Pujol, 2009; Nestler, 2013) as well as to accept poorly specified models (Nestler, 2013).

As measures of fit, the χ^2 statistics, a relative model fit comparative fit index (CFI) measure and a approximate model fit root mean square error of approximation (RMSEA) measure are reported. They are considered to be standard, most often they are presented in publications reporting the CFA results (McDonald and Ho, 2002). The χ^2 statistics should be as low as possible with as many degrees of freedom for a good fit as possible. With the significance above 0.05 a model is considered as well fit (for large samples ($n > 200$) the level of validity is usually significant). Therefore, fit evaluation should be guided by the RMSEA and CFI measures. The RMSEA lower than 0.05 is

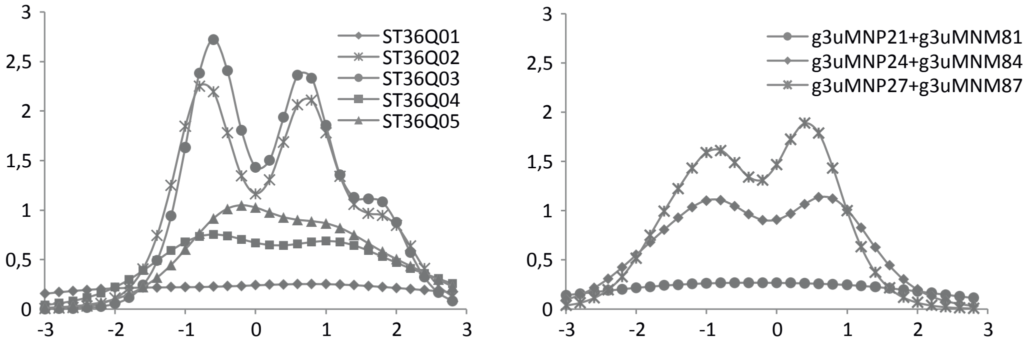

Figure 1A. Item information function for the scale used in the PISA 2009 study and its adaptation in the form of "teacher authority/classroom management" scales including the Polish language and maths jointly.

considered to be an indicator of a good fit and a value in the range of 0.05–0.08 as an acceptable fit. 90%confidence interval for the RMSEA should not contain values > 0.8. The CFI measure greater than 0.9 indicates a good fit (McDonald and Ho, 2002).

Despite some differences in fitting scales for variables referring to the Polish language and maths lessons, for the sake of clarity of interpretation, the same scale for the Polish language and maths was chosen. The "teacher authority/classroom management" scale indicators were answers to the following questions[5]:

A. "The teacher has to wait a long time for the students to quieten down."[6] [mat: 0.820 *0.007*; pol: 0.743 *0.010*], (variable name format: g3uMNN81, for pol: g3uMNP21);
B. "Students don't listen to what the teacher says". [mat: 0.818 *0.007*; pol: 0.796 *0.010*], (format: g3uMNN84, for pol: g3uMNP24);
C. "Students in class are occupied with their own matters". [mat: 0.865 *0.006*; pol: 0.792 *0.010*], (format: g3uMNN87, for pol: g3uMNP27).

With three indicators and one factor, the scales in the definition fit the data perfectly. The higher the scale values, the more "teacher authority/classroom management". This scale was adapted from the OECD PISA student questionnaire in 2009. The questionnaire is available for download at pisa2009.acer.edu.au/downloads.php. The original scale consists of the following items:

a) Students don't listen to what the teacher says (ST36Q01);
b) There is noise and disorder (ST36Q02);
c) The teacher has to wait a long time for the students to quieten down (ST36Q03);
d) Students cannot work well (ST36Q04);
e) Students don't start working for a long time after the lesson begins (ST36Q05).

The original scale has the following characteristics: n = 4892; χ^2 = 450.549; df = 5; p = 0.000; RMSEA = 0.135 (0.90 CI = 0.125 to 0.146; p = 0.000; CFI = 0.985. Cronbach's α is 0.878. Cronbach's α for the adapted scale consisting of 3 items for the Polish language is 0.769 and 0.821 for maths.

Figure 1A shows the scale characteristics. The data were fitted to the *Graded Response Model* (GRM) (Samejima, 1969). The charts present item information functions (IIF).

[5] The pupils responded according to the four-point scale: 1 – during each class; 2 – during most classes; 3 – sometimes; 4 – never or almost never. Factor loadings are given in square brackets, standard errors are given in italics.

[6] Here and elsewhere factor loadings, standard errors, model fit statistics refer to Polish version of the items and scales. Translation proposed here has been conducted only for purpose of this article and has not been field tested despite the items taken from PISA questionnaire, which have their English version.

The items g3uMNP21 and g3uMNM81; for ST36Q01 is g3uMNP24 and g3uMNM84; for ST36Q02 is g3uMNP27 and g3uMNM87 are counterparts of the item ST36Q03 of the original scale.

The "active teaching style" scale is composed of four sub-scales, the indicators of which were established according to answers to the following questions:

A. Sub-scale "individualised learning" [mat: 0.863 *0.010*; pol: 0.924 *0.011*]:
(A1) "The teacher assigns students to different tasks, taking into account their academic achievement." [mat: 0.804 *0.009*; pol: 0.756 *0.010*];
(A2) "The teacher assigns students to different tasks, taking into account their interests." [mat: 0.832 *0.009*; pol: 0.797 *0.010*].
B. Sub-scale: "multimedia use" mat: 0.807 *0.012*; pol: 0.840 *0.013*]:
(B1) "The teacher encourages the use of additional materials, such as articles, information from the internet." [mat: 0.864 *0.010*; pol: 0.44 *0.012*];
(B2) "The teacher uses multimedia materials, educational programs." [mat: 0.701 *0.011*; pol: 0.650 *0.013*];
C. Sub-scale: "revision of the material" [mat: 0.862 *0.010*; pol: 0.880 *0.014*]:
(C1) "At the end of the lesson, the teacher summarises the key topics." [mat: 0.824 *0.009*; pol: 0.724 *0.012*];
(C2) "At the beginning of the lesson, the teacher summarises what we have done in previous classes." [mat: 0.755 *0.009*; pol: 0.657 *0.012*].
D. Sub-scale: "stimulation of group work" [mat: 0.782 *0.013*; pol: 0.699 *0.015*]:
(D1) "The teacher divides the students into groups, in which students work together."
[mat: 0.770 *0.011*; pol: 0.675 *0.013*];
(D2) „Together students create posters related to the topics covered in class." [mat: 0.877 *0.013*; pol: 0.800 *0.014*].

For maths the scale has the following characteristics: mat: $n = 5071$; $\chi^2 = 232.110$; $df = 16$; $p = 0.000$; RMSEA = 0.052 (0.90 CI = 0.046 do 0.058); $p = 0.314$); CFI = 0.983. For the Polish language: pol: $n = 5069$; $\chi^2 = 241.969$; $df = 16$; $p = 0.000$; RMSEA = 0.053 (0.90 CI = 0.047 to 0.059); $p = 0.208$); CFI = 0.966. Fit to the data is acceptable. To facilitate interpretation, the scale values have been multiplied by -1. The higher the scale value, the more the "activating teaching style".

Indicators of the "helpful teacher" scale were the answers to the following questions:

A. "The teacher shows interest in learning progress of each student." [Mat: 0.709 *0.009*; pol: 0.660 *0.010*];
B. "The teacher explains until all students understand." [Mat: 0.871 *0.006*; pol: 0.796 *0.007*];
C. "The teacher is willing to help students learn." [Mat: 0.896 *0.006*; pol: 0.874 *0.007*];
D. "The classes are carefully planned." [Mat: 0.652 *0.010*; pol: 0.650 *0.010*];
E. "The teacher encourages students to ask questions in class about the material covered." [Mat: 0.669 *0.010*; pol: 0.603 *0.011*].

For maths the scale has the following characteristics: $n = 5072$; $\chi^2 = 97.626$; $df = 5$; $p = 0.000$; RMSEA = 0.060 (0.90 CI = 0.050 to 0.071); $p = 0.045$; CFI = 0.994. For the Polish language: $n = 5071$; $\chi^2 = 86.771$; $df = 5$; $p = 0.000$; RMSEA = 0.057 (0,90 CI = 0.047 to 0.068); $p = 0.131$; CFI = 0.991. Fit to the data is acceptable. To facilitate interpretation, the scale values have been multiplied by -1. The higher the scale values, the more "helpful" the teacher.

Appendix 2 – hierarchical models

Hierarchical models were calculated with the HLM7 software. Only random intercepts models were used. When being introduced to the models, all variables were "non-centred". Of course, these variables which in their original form are expressed on the scale *Z*, have a mean equal to 0 and standard deviation equal to 1.

1MAT: ZIRTMAT = γ_{000} + γ_{100}*ZSP3PL + γ_{200}*EWD_PLEC + γ_{300}*DYS + γ_{400}*RAVEN + γ_{500}*HEDULL + γ_{600}*HOMEPOS + γ_{700}*WIELRODZ + *r0* + *u00* + *e*
2MAT: ZIRTMAT = γ_{000} + γ_{001}*LOK + γ_{002}*YOYBEZ + γ_{010}*ZSPRAWD2 + γ_{100}*ZSP3PL + γ_{200}*EWD_PLEC + γ_{300}*DYS + γ_{400}*RAVEN + γ_{500}*HEDULL + γ_{600}*HOMEPOS + γ_{700}*WIELRODZ + *r0* + *u00* + *e*
3MAT: ZIRTMAT = γ_{000} + γ_{001}*LOK + γ_{002}*YOYBEZ + γ_{010}*ZSPRAWD2 + γ_{020}*ACTIVM5 + γ_{100}*ZSP3PL + γ_{200}*EWD_PLEC + γ_{300}*DYS + γ_{400}*RAVEN + γ_{500}*HEDULL + γ_{600}*HOMEPOS + γ_{700}*WIELRODZ + *r0* + *u00* + *e*
4MAT: ZIRTMAT = γ_{000} + γ_{001}*LOK + γ_{002}*YOYBEZ + γ_{010}*ZSPRAWD2 + γ_{020}*DYSM + γ_{100}*ZSP3PL + γ_{200}*E-WD_PLEC + γ_{300}*DYS + γ_{400}*RAVEN + γ_{500}*HEDULL + γ_{600}*HOMEPOS + γ_{700}*WIELRODZ + *r0* + *u00* + *e*
1POL: ZIRTPOL = γ_{000} + γ_{100}*ZSP3PL + γ_{200}*EWD_PLEC + γ_{300}*RAVEN + γ_{400}*HEDULL + γ_{500}*HOMEPOS + γ_{600}*WIELRODZ + *r0* + *u00* + *e*
2POL: ZIRTPOL = γ_{000} + γ_{001}*LOK + γ_{100}*ZSP3PL + γ_{200}*EWD_PLEC + γ_{300}*RAVEN + γ_{400}*HEDULL + γ_{500}*HOMEPOS + γ_{600}*WIELRODZ + *r0* + *u00* + *e*
3POL: ZIRTPOL = γ_{000} + γ_{001}*LOK + γ_{010}*POMOCNP + γ_{100}*ZSP3PL + γ_{200}*EWD_PLEC + γ_{300}*RAVEN + γ_{400}*HEDULL + γ_{500}*HOMEPOS + γ_{600}*WIELRODZ + *r0* + *u00* + *e*
4POL: ZIRTPOL = γ_{000} + γ_{001}*LOK + γ_{010}*DYSP + γ_{100}*ZSP3PL + γ_{200}*EWD_PLEC + γ_{300}*RAVEN + γ_{400}*HEDULL + γ_{500}*HOMEPOS + γ_{600}*WIELRODZ + *r0* + *u00* + *e*

Figure 2A. Formulas for the hierarchical models reported in this article.