European University Institute
Robert Schuman Centre for Advanced Studies

EUI Working Papers

RSCAS 2008/06

EUROPEAN FORUM 2006/2007

Implementing Value-Added Models of School Assessment

Maciej Jakubowski

EUROPEAN UNIVERSITY INSTITUTE, FLORENCE
ROBERT SCHUMAN CENTRE FOR ADVANCED STUDIES
EUROPEAN FORUM 2006/2007

Implementing Value-Added Models of School Assessment

MACIEJ JAKUBOWSKI

EUI Working Paper **RSCAS** 2008/06

ISSN 1028-3625

Printed in Italy in February 2008
European University Institute
Badia Fiesolana
I – 50014 San Domenico di Fiesole (FI)
Italy
http://www.eui.eu/RSCAS/Publications/
http://cadmus.eui.eu

Robert Schuman Centre for Advanced Studies

The Robert Schuman Centre for Advanced Studies (RSCAS), directed by Stefano Bartolini since September 2006, is home to a large post-doctoral programme. Created in 1992, it aims to develop inter-disciplinary and comparative research and to promote work on the major issues facing the process of integration and European society.

The Centre hosts major research programmes and projects, and a range of working groups and ad hoc initiatives. The research agenda is organised around a set of core themes and is continuously evolving, reflecting the changing agenda of European integration and the expanding membership of the European Union.

Details of this and the other research of the Centre can be found on:
http://www.eui.eu/RSCAS/Research/

Research publications take the form of Working Papers, Policy Papers, Distinguished Lectures and books. Most of these are also available on the RSCAS website:
http://www.eui.eu/RSCAS/Publications/

European Forum

The European Forum brings together scholars from the EUI and other research institutions to carry out comparative and interdisciplinary research on a specific annually chosen topic under the guidance of the Director of the RSCAS and the Forum's annual scientific director(s).

The 2006-2007 European Forum, 'Assessing the Quality of Education and its Relationships with Inequality in European and Other Modern Societies', was directed by Prof. Jaap Dronkers, a sociologist in the Social and Political Sciences Department at the EUI, and a leading expert on cross-country comparative social research.

The aim of that European Forum was to explore the use of the PISA data-sets, but also other relevant cross-national data-sets, to provide answers to scientific and policy questions on education and its relationships to various forms of inequality in European societies (including economic, legal and historical dimensions).

For further information:

E-mail: forinfo@eui.eu

http://www.eui.eu/RSCAS/Research/EuropeanForum/Index.shtml

Abstract

This paper considers value-added models of school assessment and their implementation in Poland. Value-added estimates can be very helpful for schools and policy makers who need a reliable way to control teaching effectiveness, or for parents who need information about school quality in their area. However, their usefulness depends on several statistical issues and specific decisions made during implementation. The paper discusses several value-added models and describes details of the solution implemented in Poland. Statistical problems are discussed according to their policy relevance. It is shown that what bothers statisticians is less important in practice than several problems encountered when one wants to apply these models to a policy relevant context. Problems of proper regression specification, omitted variables bias, and measurement error are discussed, but the ways value-added estimates could be published and used as policy evaluation tools are also presented. All this problems are discussed from a practical point of view using three years of experience in implementation of these methods in Poland.

Keywords

education, school assessment, school effectiveness, value-added models

JEL: I21, I28, J24

I. Introduction*

The education system in Poland has changed heavily during the last 17 years. Decentralization and market-oriented reforms have been part of an overall public sector transformation. One of the most profound changes has been the introduction of an external examination system. From 2002 each student at primary and lower secondary school has to write an exam, which is the same across the whole country. From 2005 students at the upper secondary school level also have to write the '*Matura*,' which serves as a university entrance exam. A growing interest in the quality of public schools and examination results can be observed. This is due to an increasing awareness of the value of a good education in a market economy.

There is a strong conviction among education experts and officials in Poland that unprocessed external examination results are of little value when one wants to assess the quality of teaching or school effectiveness. Thus, in 2005 a group of experts was established to analyze, in close collaboration with the Central Examination Board, the possibilities of value-added assessment in Poland. This report is based on two years of research and policy experience of the expert group. During that time the group proposed a value-added model to assess the effectiveness of lower secondary school. This was implemented after many discussions with several groups of stakeholders, including lower secondary school principals. In 2008 first value-added estimates will be made publicly available. Thus, we still need to wait to see how the new policy will affect school environments. Nevertheless, it seems interesting to discuss the implementation of value-added models in Poland. What were the pros and cons and how they interplayed with the Polish school system. The paper focuses on statistical issues but confronts them with policy objectives. This is rarely the case that these problems are discussed together, but there is no point to discuss value-added models only from one of these perspectives. In this case statistics have to go adjust to policy and policy-makers have to confront limitations, but also the rigor of statistical procedures.

This report is organized as follows. In section II country background is given. Section III explains policy objectives for value-added in Poland. Section IV briefly describes available data. Section V contains a description of the value-added model for lower secondary schools proposed in Poland and section VI compares this model with other methods of value-added assessment. Section VII describes how measurement error affects value-added estimates. Section VIII proposed methods of public dissemination of value-added results are described. In section IX policy interpretation and critical assessment of value-added methods in the Polish context are presented. Section X gives an example of how value-added models could be used in policy evaluation, in this case to assess the impact of decentralized expenditures on teaching quality. Section IX summarizes and gives proposals for future developments.

II. The examination system in Poland

There are two levels of compulsory comprehensive education in Poland. The first is a six-year primary school together with preparatory 'zero' classes, which were made obligatory in 2004. The second is a three-year lower secondary school called '*gimnazjum*.' In 2002, for the first time, all students who were to finish primary or lower secondary school had to take an exam conducted by governmental agencies called Regional Examination Boards. Exams were prepared and supervised by a Central Examination Board and were uniform across the country. Similar exams are repeated each year with basic characteristics unchanged until now.

* This project was conducted during the author's stay at the RSCAS EUI generously supported by the Foundation for Polish Science. Author would like to thank Central Examination Board (CKE) in Poland for invaluable support.

The exam at the end of primary school is called *sprawdzian* (competence test) and its results are neither publicly available at the individual level nor can be officially used to help lower secondary schools' principals in selection among candidates[1]. *Sprawdzian* is a fairly simple multi-subject test aimed at assessing students' ability to learn in the lower secondary school. It is reported on a 0-40 points non-standardized scale. It is impossible not to pass this exam and even students with zero points have to go to the lower secondary school if their school grades at the end of the year were sufficient. It is assumed that the role of *sprawdzian* is mainly informative.

The exam at the end of lower secondary school is called *egzamin gimnazjalny* and consists of two parts: one aimed at measuring the level of knowledge in mathematics and science, and the second aimed at measuring knowledge in humanities. Results from both parts are reported on 0-50 point non-standardized scales. Unlike *sprawdzian* the lower secondary school exam can be used in admission decisions by upper secondary schools' principals which is quite important in urban areas with the biggest number of schools where principals can choose freely among candidates from other districts.

In 2005, students ending the three-year *Liceum* and, from 2006 all students finishing the four-year *Technikum* had the possibility to pass the new *Matura* exam which starting from this point is also conducted by Examination Boards and was the same for all students in Poland. The new *Matura* should be treated by universities as an entrance exam and is separately conducted for more than 60 subjects. The other type of upper secondary school is a vocational school with a special external exam ("vocational exam"), which in theory should serve as a document for future employers.

Exams at the end of upper secondary schools were introduced recently and produce results that are hard to interpret due to non-standardized subject scores and the non-comprehensiveness of these schools which creates sample selection bias. Exams at the end of both kinds of comprehensive schools (primary and lower secondary) are different in this regard. These exams are multi-subject, obligatory, easier to interpret and seem more valid as a basis of value-added models.

Each year exam results are published at the country and regional levels. Individual results are available to students, parents, teachers and school principals. All results are reported on a non-transformed raw point scale and on the 9-point stanine scale, which locates a given school in the distribution of all scores in the country. Thus, each school receives an average score and some means of comparison with other schools in the country, but datasets with the results of all schools in Poland are not publicly available. A few regional examination boards have decided to publish school results on their web sites and some newspapers had published schools' scores from their area, but the Central Examination Board still does not make school results freely available for comparisons at the country level. During the last year a website was created when one can see results of all schools in Poland, but separately. Thus, it is possible to create league-tables if one wants to spend several ours downloading the results, however, we don't know any attempts to do this for whole country or even one of the regions. There is still room for public involvement and regulation to protect schools from improper comparisons.

III. Policy motivation for the development of value-added models

Poland is a diverse country with important regional variation in the cultural, social and economic background of the school systems[2]. Results of external exams confirmed non-negligible disparities between the achievements of rural and urban students and between students from different historical parts of the country. Some experts claim that decentralization and the introduction of market forces could further widen achievement gaps and that between-school differences are of growing importance.

1 Public primary and lower secondary schools principals have to accept all students from their school district, however, they can choose among candidates from other districts.

2 More up to date information about school system in Poland could be found in: O'Brien, Paczyński (2006).

In this context, test results are seen not only as a kind of assessment of student achievement but also as a tool for school system supervision. Some important questions of correlation between school effectiveness, socio-economic background and financing have to be answered. Increasing awareness of the need for wide scale quantitative measurement of teaching quality is one the fundamental issues underlying the development of value-added models.

The other issue, which in fact is seen by many practitioners as a main reason for value-added assessment, is the problem of misinterpretation of exam results by parents, teachers, school principals and local government who is the owner of 98% of schools in Poland. Mean results of schools are very often erroneously treated as measures of teaching quality. They serve as a basis for school comparison and in some places are utilized for accountability purposes. In effect, teachers and school principals became frustrated at new examination schemes, especially in rural areas and poorer neighbourhoods. From this point of view a claim that introduction of external examination and publication of test results could increase discrepancies between schools seems justified, because good teachers can more frequently choose to work in schools with "good" results.

For many experts and stakeholders it has become obvious that a new way of analyzing exam results is needed. Clearly, some measures of school effectiveness should be developed. Such measures are needed at the country level to quantitatively control between-school differences in teaching quality and at the local level where local governments and parents do not have comparable information about school effects in their area. The lack of any measures of this type forces improper interpretations of school results. At the same time a few regional boards, *Kuratoria* (governmental bodies responsible for supervision of schools) and even school principals have tried to develop their own measures of teaching effectiveness based on very simple notions of value-added (e.g. by taking a difference of primary and lower secondary school's average exam score or even stanine score). These attempts were in most cases of poor quality and it was not possible to use them at the country level. The need for the development of proper value-added assessment models at the country level was quite obvious.

In 2005 the Central Examination Board established a group of experts to research models of value-added assessment that could be implemented in Poland. In 2005 results from the lower secondary school exam were available for the cohort of students who sat the primary school exam (during the first year of examination in 2002). Thus, value-added analysis was for the first time possible at the country level. After one year of research the group proposed a model of value-added for lower secondary schools. This model was discussed with a wide group of stakeholders (representatives of *Kuratoria*, teachers' professional development assistants, school principals) and experts. At the same time some qualitative research on the accuracy of the new method in measuring school effectiveness was conducted, part of which is presented in this paper. The expert group did not propose a model for upper secondary schools for reasons mentioned above, i.e. non-comprehensiveness and difficulties in result interpretation. The value-added model for primary schools is not possible because there is no exam at the beginning or during the schooling period.

The objective of the research group was to compare different value-added models from the theoretical point of view and to analyze them in the context of implementation and usefulness for the public school system in Poland. We assumed that raw test results reported to schools should be accompanied by measures of school effectiveness to limit misinterpretation. We also assumed that value-added methods should be able to serve as a monitoring tool for school supervising bodies, Examination Boards and the Ministry of Education. We also emphasized the point that value-added measures could be useful tools at the school level, so we planned to develop methods of value-added analysis of student results which school principals and other local actors could find helpful.

Part of results presented in the paper were conducted on the data for a first cohort for which value-added analysis was possible (taking 2002 primary school exam and 2005 lower secondary school exam). Now we have also results for the 2003/2006 and 2004/2007. The correlation between two-year estimates of value-added and analysis of its volatility was researched elsewhere (Jakubowski, 2007c).

It is enough to mention that value-added scores are stable enough to use them as indicators of school quality. Thus, in 2008 value-added information will be published and we will se how school system reacted to them. Now, we would like to report findings from more than two years of research on value-added models for lower secondary schools in Poland with strong emphasis on their policy relevance.

IV. Data considerations

To conduct a value-added assessment one obviously needs test results from at least two separate points in time. In Poland there are three public external exams: at the end of primary school, at the end of lower secondary school and at the end of upper secondary school. This means that value-added assessment is possible only for: (a) lower secondary schools based on results for primary school and lower secondary school exams; (b) upper secondary schools based on results for lower secondary schools and upper secondary schools. It was explained earlier why value-added assessment of upper secondary schools does not seem possible at this time. Decisions have to be made as to which subjects of the exam should serve as the basis for school assessment but it is not even clear what will be the obligatory content of this exam in the nearest future (e.g. if the math exam will be obligatory) which makes any proposal of value-added temporary.

Hence we concentrated on value-added assessment for lower secondary schools. To start with value-added analysis of any kind, intake exam scores need to be connected with final exam scores for each student. Examination boards' databases were not projected to automatically fulfil such needs. No single ID number for students exists and the results from different exams had to be matched individually using date and place of birth, gender and name. Matching was done by regional boards because the Central Examination Boards do not store individual characteristics of students and some information was surely lost because of some student moved between regions. The linking process was not successful for less than 10% of students. Thus, the early finding of the research group was that the data collection process has to be done centrally if value-added assessments are to be delivered on time. In addition, a common student identification system has to be implemented if value-added is to be used for all schools and students.

Finally, about 90% of student results from 2002's *sprawdzian* (exam at the end of primary school) and from 2005's *egzamin gimnazjalny* (exam at the end of lower secondary school) were available for analysis. For each student the team had an exam score and some typical characteristics: gender, date and place of birth, school and region ID, and dummy variables indicating whether the student had dyslexia during the time of the exam[3]. After the first year of research additional data for the 2003/2006 cohort were available and at the end 2007 data for 2004/2007 cohort were merged. However, they confirmed earlier findings and were used mainly for volatility study we already mentioned. We did have data for upper secondary schools, but for reasons mentioned earlier we finally decided that they cannot be used for valued-added school evaluation.

For privacy reasons data collected by examination boards do not contain characteristics of student socio-economic background. Therefore, a full analysis controlling for student SES is not possible. There are some additional datasets on schools which are collected by the Ministry of Education through the Educational Information System (SIO). SIO database contains very detailed data on school equipment and organization as well as aggregated school-level teacher characteristics. However, data

3 Students with dyslexia write similar test, however, they are graded differently for particular questions and their writing time is extended. In some regions parents believe that it is easier to score higher for students writing the 'dyslexia' version of the exam and they do their best to qualify their children as dyslexic. In some schools the percentage of students officially classified as 'with dyslexia' reached 60-70%. Students with dyslexia score less on the *sprawdzian* and higher on both parts of *egzamin gimnazjalny*. Having in mind that in some schools the percentage of dyslexic students is so high, it is obvious that one needs to control for dyslexia to obtain unbiased value-added estimates.

on SES are not collected. Discussions with the Ministry of Education and Examination Boards revealed that collecting such data is against existing law and will not be possible in the nearest future.

A typical source of information in Poland is the Central Statistical Office, which collects detailed data not only on schools but also on labour markets, adult education and other areas. Unfortunately, these data are aggregated to the local government level and are useless as control variables in school-level analysis. For example, the educational attainment of adults or labour market characteristics have one value for cities in which might be located more than 100 schools to be assessed.

Thus, in the nearest future models of value-added assessment for all Polish schools cannot utilize SES characteristics. However, data collected during several research projects conducted by Examination Board on random samples of students to investigate determinants of exam results give the possibility to test whether including SES variables in the value-added models could significantly change obtained estimates. We discuss main findings from such research.

V. Value-added model proposed for lower secondary school assessment

In this section a value-added model proposed for lower secondary schools in Poland is described. This model was chosen after careful analysis of several different value-added methods. It was assumed that the chosen model should be:

- *theoretically valid but fairly simple* – method underpinnings should be explicable to people with a basic knowledge of statistics and proper interpretation of estimates should be understandable for all recipients, including parents as well;
- *neutral* – probability of positive or negative value-added assessment of a school should be independent of its students intake scores;
- *easy to implement* – value-added models should be easy to adapt by local Examination Boards or *Kuratoria*; re-analysis of value-added results at the local level should be possible in order to support school curricula and teaching methods development.

The results of the comparative analysis of several methods are presented in section VI below. Here details of the proposed method are described.

The proposed method is based on a simple linear regression model where individual student-level lower secondary school exam scores are regressed on scores from the exam conducted at the end of primary school. Let x_i be the exam score of i-th student at the end of primary school and y_{ij} be the exam score for j-th subject of i-th student at the end of lower secondary school then:

$$y_{ij} = \beta_0 + \beta_1 x_i + \mathbf{R}_i'\boldsymbol{\beta}_2 + \mathbf{P}_s'\boldsymbol{\beta}_3 + \varepsilon_{ij} \qquad (1)$$

is a linear regression model where β's are parameters to be estimated, vectors $\mathbf{R}_i$ and $\mathbf{P}_s$ contain characteristics at the student and school level respectively and ε_{ij} is the residual error term. Residuals from estimated regression (1) are of interest here. Value-added measure for a s-th school is a mean residual of students from this school:

$$VA_{js} = \frac{1}{n_s}\sum_{i \in S} \varepsilon_{ij} = \frac{1}{n_s}\sum_{i \in S}\left(y_{ij} - \hat{y}_{ij}\right) \qquad (2)$$

where S is a set containing all students of s-th school, n_s is a number of students in that school, and $\hat{y}_{ij}$ is an estimated linear regression prediction of upper-level j-th subject exam scores.

In practice a proper specification of regression model (1) is the main difficulty in this method. As we said earlier, relevant school level characteristics (vector $\mathbf{P}_s$) are not available in the Polish case and

in the final model individual level characteristics (vector $\mathbf{R}_i$) contain only dummies for gender and dyslexia. Clearly, a more complex set of control variables is needed to assess school teaching quality independently of economic and social background. However, when one needs a measure of the overall relative effect of a particular school on student score gain, then control variables are not necessarily needed. Such overall effect measures are more proper for parents than for supervisory bodies. We will return to interpretation and valid use of value-added estimates in the section IX.

Another issue is a proper specification of the functional relationship between exam scores. In the context of value-added measurement such specifications have special meaning. For example, the quadratic relationship between exam scores can be interpreted as a sign of a differential score gains effect for students with different intake score levels. In other words, it assumes that better students gain more. However, such non-linear relationships can also be observed in the case of non-normally distributed exam scores. Additionally, it is possible that if better students collectively go to schools of better quality then it can produce such non-linearities. We cannot empirically distinguish between these three cases. In the Polish model we used unprocessed exam scores which were heavily skewed for the primary school exam and for the math-science part of the lower secondary school exam. In effect, the relationship between scores from these exams was non-linear, but no one can be sure that this solely due to skewed scores distribution. It could be also true that other two causes are valid.

To resolve this issue we decided not to assume a priori any relation between exam scores and that we should choose among specifications that give conditional mean residuals close to zero for every level of intake exam score. Put differently, the chosen functional specification should give means of ε_{ij} as close to zero as possible for every level of x. This assured that schools will not gain or lose because of the mean level of their student intake score. Clearly, we wanted to avoid value-added artefacts produced by improper model specification. However, based on this assumption we rejected the possibility that schools could systematically differ in their effectiveness with regard to average student knowledge and ability levels or if this is the case then one should interpret value-added scores as relative to other schools with similar student intake scores.

Final specifications for 2002/2005 data were described in the official reports (Dolata, 2006; Jakubowski, 2006a, 2006b). To solve the non-linearity problem we estimated linear spline regressions. Additionally, we added squared individual intake scores to allow a quadratic relationship for higher values of primary school exam results. It shows that in 2006 and 2007 specifications were very similar. Thus, it seems that there are related to the nature of subject-specific measurement imperfections produced by still developing examination tests in Poland.

After estimating value-added scores, one can further explore data to analyze school effectiveness in more detail. For example, we assumed that value-added is measured as the mean of residuals for a particular school but one can use other measures of central tendency. Mean is a non-robust statistic and for small schools could be misleading. In the presence of outliers (i.e. residuals of students who for any reason gave back a blank test sheet on the primary school exam but obtained a high score from the upper-level exam) robust measures of central tendency are more advisable, for example median or trimmed-mean. Median can also be used to test whether distribution of residuals is skewed which can be the case when a school is less effective with a specific group of students (e.g. one class, boys or a particular teacher's students). Other simple statistics such as standard deviation can be used and this deeper analysis can easily be done by trained teachers or other local actors if the proposed simple value-added method is employed. Simply, one can use predicted values to compute regression residuals and then use them in analysis of any kind.

VI. Comparison with different value-added models

In this section the results of comparative research on different value-added models are presented. Estimates obtained by employing the statistical model described in section V (called *the basic model*)

are treated as a base-line for comparisons. For every analyzed method a correlation coefficient of estimates obtained by this method and by the basic model is presented. We start with a comparison of different specifications of the basic model. Then we compare the basic model with other methods of value-added assessment. Next, we test whether extending the set of control variables to include SES characteristics alter results. Finally, we show discrepancies between subject-specific value-added estimates.

VI.1. Basic model specification

The problems of proper specification of regression equation (1) in the context of Polish exams have already been discussed. In this subsection we compare the effects of: (a) differently specified functional form of intake scores; (b) different measures of central tendency of residuals; and (c) inclusion of control variables. In table 1 below, correlation coefficients between estimates of school value-added obtained by different specifications of the basic model are presented. Each model was run separately with scores from humanities and math-science parts of the exam as well as on the total exam score. Regressions were estimated on a sample of 483,692 students who finished primary school in 2002 and left lower secondary school in 2005. Students in the sample are from 6256 schools – almost all lower secondary schools in Poland.

Table 1. Correlation between value-added estimates obtained through different specifications of the basic model.

	humanities	***math-science***	***sum***
linear	0.999	0.967	0.986
quadratic	0.999	0.999	1.000
categorical	0.999	0.999	1.000
trimmed mean	0.997	0.997	0.997
median	0.971	0.968	0.974
no control	0.994	0.996	0.998

The first three rows show the correlation between the basic spline regression model and similar models, which differ only by functional form of primary school exam scores (linear or quadratic). 'Categorical' means that primary school exam scores were treated as measured on the ordinal scale so dummies for each value of the variable were included separately into the regression equation. Thus, in this case no functional relationship between exam scores was assumed. We see that in the case of the math-science part and total exam score one should consider non-linear relationship between exam scores but different specifications (spline regression, quadratics, categorical) give very similar results.

The next two rows present the correlation between estimates from the basic model, where the mean of residuals was used to calculate school value-added, and estimates based on different central tendency statistics. 10% trimmed mean and median were used to check whether they differently estimate the central tendencies of schools' residuals. It seems that using the trimmed-mean gives similar results to using the mean. Employing the median changes estimates slightly, however, the correlation coefficient is still high.

The last row shows the correlation between estimates from the basic model where dummies for gender and dyslexia were used and a model without any control variables. Not surprisingly, correlation

is very strong here, because of the lack of important individual characteristics. We return to the issue of including SES variables below.

VI.2. Comparison of different value-added methods

In this subsection the basic model is compared with more theoretically valid models. First, we estimated school value-added as fixed effects in a regression model. In some circumstances value-added measures obtained by these two methods can differ (see Ladd, Walsh, 2002). Many scholars believe that one should apply a multilevel model to estimate value-added as a school random effect (or random intercept)[4]. We also applied this model to see whether it makes a difference.

To compare all methods we applied a model where all variables were defined as in the basic model except a school fixed or random intercept. Namely, we estimated regression defined by the equation given below:

$$y_{ijs} = \alpha + \beta_1 x_{ij} + \mathbf{R}_i'\boldsymbol{\beta}_\mathbf{2} + \mathbf{P}_s'\boldsymbol{\beta}_\mathbf{3} + u_s + \varepsilon_{ijs} \qquad (3)$$

where in addition to the equation (1) the term u_s was added. We estimated this regression assuming that u_s is a school fixed effect or that u_s is a school random effect. In both cases predicted values of these effects were used as the value-added estimates of a particular school. In the case of random effects we estimated this model through GLS and MLE to see whether there is any practically visible difference.

We also used random slope models to see how estimates of school effectiveness depend on the intake scores of students. Thus, we estimated two regressions where intake scores were centred around the 10th and 90th percentiles of their distribution. The random slope model is given by the equation below:

$$y_{ijs} = \alpha + (\beta_1 + \upsilon_s)\tilde{x}_{ij} + \mathbf{R}_i'\boldsymbol{\beta}_\mathbf{2} + \mathbf{P}_s'\boldsymbol{\beta}_\mathbf{3} + u_s + \varepsilon_{ijs} \qquad (4)$$

where in addition to the equation (3) the term υ_s was added and intake scores $\tilde{x}_{ij}$ were centred around the relevant percentiles. As usual we assumed in all regressions that random effects are normally distributed. In the case of random slope regressions we additionally assumed that the covariance between slope and intercept is non-zero allowing any interaction between those two effects.

In table 2 correlation between estimates obtained from the basic model with those obtained from the fixed effect model, and random effects models estimated through GLS and MLE (empirical Bayes predictions) are presented[5]. We see that both random effects estimation methods give similar results which are in fact quite different from the basic model and fixed effects estimates. The discrepancy in estimates between fixed and random effects models is due to "shrinkage" which heavily changes estimates for smaller schools[6]. When schools with less than 10 exam scores were excluded from calculations, correlation coefficients were much closer to 1. Thus, shrinkage could be a way to avoid false conjectures about smaller school effectiveness without arbitrarily defining a school size threshold for which value-added analysis is possible. One should note that the number of relatively small schools is not negligible in Poland. In the 2002/2005 sample 7% of schools had less than 10 students

4 See McCaffrey et. al. (2005) for a thorough discussion of more and less advanced value-added models.

5 All calculations were done using Stata statistical package (with -xtmixed- or -xtreg- procedures). See book by Rabe-Hesketh and Skrondal (2005) or Stata reference manual for detailed description of employed estimation methods.

6 Shrinkage depends on the number of observations within a school and on the variability of within school scores (see Raudenbush, Bryk, 2002, for details).

writing the exam and 15% had less than 20. Thus, the question of how to deal with such schools in the value-added assessment is of importance. Shrinking estimates seems a justified solution.

In the last two rows of table 2 correlation between the basic model estimates and value-added measures from the model with random intercept and slope are presented. Such measures could be used to compare schools' effects for students of different achievement levels. We see that correlation coefficients are around 0.9 here. Correlation between estimates obtained by the random intercept model and the random slope/intercept model was from 0.94 (for math-science) to 0.98 (for humanities). Finally, the correlation between value-added calculated for low-achievers and high-achievers (10^{th} and 90^{th} percentiles of intake score distribution respectively) range from 0.84 (for math-science) to 0.93 (for humanities).

Table 2. Correlation between value-added estimates obtained from the basic model and obtained from different methods

	humanities	***math-science***	***sum***
fixed effects	0.999	0.998	0.999
random effects (GLS)	0.927	0.954	0.944
empirical Bayesian random effects	0.918	0.952	0.939
random slopes (10^{th} percentile)	0.900	0.896	0.910
random slopes (90^{th} percentile)	0.895	0.916	0.911

Hence, there is some evidence that, at least for some schools, value-added measures depend on the achievement levels of their students. If school effectiveness is different for sub-groups of students, then value-added methods should account for this not only for accountability reasons, but also to inform schools more deeply for self-assessment and development purposes. However, random slope models are fragile to any discrepancies in distribution of scores and non-linearities which are present in our datasets (Goldstein, 1997). Thus, these results should be treated as preliminary.

VI.3. Controlling for SES

As stated earlier, SES variables are not available for analysis at the country level. However, smaller scale random sample representative studies could be used to test whether including SES variables will noticeably change obtained value-added estimates. Here, we explore a dataset collected in 2006 as part of the research program of the Central Examination Board. The sample consists of 1233 individuals from 81 schools for whom data on parent education and household's earnings were collected[7]. Before we turn to results it should be noted that this research substantially differs from the earlier analysis based on exam scores of the full population of students. First, not all students responded and in each school questionnaires were distributed only in one class chosen at random. Second, student intake score was self-declared on a less detailed scale (from 0 to 10 – originally 0 to 40). Third, the sample is representative for students and not for schools. Thus, a comparative study of value-added methods based on this sample can be generalized only with some caution.

To compare the effect of SES on value-added, we estimated two types of models with and without SES variables. First, the basic model with and without SES controls was estimated. Second, the

7 The whole sample contains more than 3000 observations, however, a full questionnaire was given to about 1500 students and some of these did not respond to questions about earnings (see Jakubowski, 2007a).

random intercept model with and without SES controls was estimated. Based on answers from questionnaires, two variables were constructed: (1) summarizing parent educational attainment; (2) a measure of household equivalent income per person calculated using the standard OECD scale. These variables were found to be good predictors of student achievement and were strongly correlated with other important characteristics of family and school socio-economic background (i.e. employment status and type of employment sector, age, school localization and composition).

In Table 3 below, correlation coefficients between estimates obtained by different models with and without SES controls are presented separately for different subjects and for the exam total score. Correlation between estimates from the basic model with and without SES variables varies from 0.74 (for humanities) to 0.80 (for math-science). Correlation between estimates from the random intercept model with and without SES controls varies from 0.80 (for humanities) to 0.82 (for math-science). Clearly, adding SES variables alters value-added considerably.

Table 3. Correlations between value-added estimates obtained by models without and with SES control variables

	Basic model without SES	Bayesian random effects without SES	Basic model with SES
		humanities	
Bayesian random effects without SES	0.965		
Basic model with SES	0.739	0.658	
Bayesian random effects with SES	0.781	0.805	0.844
		math-science	
Bayesian random effects without SES	0.976		
Basic model with SES	0.796	0.737	
Bayesian random effects with SES	0.802	0.821	0.877
		sum	
Bayesian random effects without SES	0.973		
Basic model with SES	0.767	0.695	
Bayesian random effects with SES	0.799	0.815	0.860

One should note that the correlation between estimates from the basic model and from the random effects model is much stronger here than in the case of the whole population of students. Thus, in the whole population the effect of including SES can be even more dramatic. More research in this area is certainly needed to understand whether omitting SES variables importantly biases value-added estimates.

VI.4. Correlation between value-added estimates for different subjects

Value-added measures for schools can shed light on their overall effectiveness as well as subject specific effects. Here the problem of separating teacher effects from school effects is an issue. We do not have data linking students to their teachers and we cannot say how deeply school effects depend on the quality of work of a particular teacher. However, we can touch the issue here by looking at

discrepancies between subject specific value-added estimates, because in Poland subjects tested in the humanities and math-science parts of the exam are taught by completely different sets of teachers.

In table 4 below we clearly see that the correlation between value-added for the humanities part and the math-science part is quite weak. This makes presenting one value-added measure for the school problematic. One could use estimates based on total exam score to say whether a school is more or less effective but the truth is that in most cases subject specific estimates considerably differ. Therefore, the choice between estimating one value-added measure for a school or just reporting both subject specific estimates remain an important question to be resolved in the future.

Table 4. Correlation between different subjects value-added estimates obtained by the basic model

	humanities	math-science
math-science	0.6003	
sum	0.8820	0.9063

VII. Impact of measurement error on the value-added estimates

The measurement error in value-added models is mainly caused by imperfect measurement during the tests. Educational tests measure achievement with limited reliability and test conditions or other circumstances could have great impact on individual scores and its variation. Moreover, even after standardization measurement errors are heteroskedastic with greater variance at the extremes of the distribution of true achievement. This is because tests are specified to accurately measure the middle of achievement distribution. Measurement errors account for a sizable fraction of the variability in scores and any systemic errors are likely to have great impact on VA estimates. To our knowledge no official value-added implementation explicitly account for measurement error or heteroskedasticity (see McCaffrey D., Lockwood J., Koretz M., Hamilton L., 2005). However, it is possible to see what impact measurement error have on a distribution of value-added scores. More specifically, we are interested in looking at the correlation between the average intake score and the difference between scores with and without measurement error correction.

The typical discussions of measurement error in econometrics describe how it affects regression coefficients. However, in our case we want to see how measurement error affects value-added estimates. We follow the approach of Ladd and Walsh (2002) using instrumental variables correction for measurement error. Namely, we estimated two-stage least squares regression with pre-test primary school scores as the instrument for the primary school final exam scores[8]. This way we corrected measurement error in intake scores and could observe what difference it makes for value-added estimates. We extend the approach of Ladd and Walsh to random effects model to see how measurement error affects value-added estimates in this case[9]. In multilevel random effects models the impact of measurement error is more complicated and less predictable (see Woodhouse, Yang, Goldstein, Rasbash, 1995, for general discussion of measurement error in multilevel models).

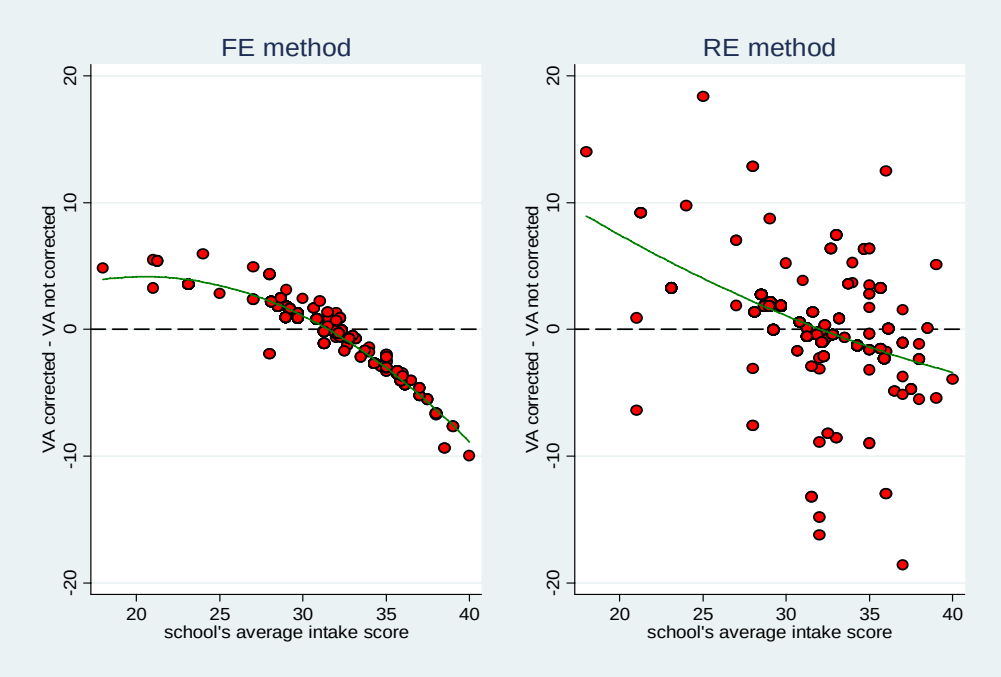
8 Pre-tests were conducted on a sample of schools to assess the test reliability in Małopolska region in Poland. Results of pre-tests were highly correlated with final exam scores and could be used as instruments.

9 More specifically we used fixed effects and random effects GLS two-stage least squares panel data models implemented in Stata statistical package in a procedure -xtivreg- (see Stata, 2007; Baltagi, 2005)

The research was done with pre-test scores and final test scores for ca. 900 students in 83 primary schools in one of the regions of Poland. While distribution of pre-test and final-test scores differ, their relation to lower secondary school exam score is very similar. For brevity we do not present regression estimates which are similar to those for whole population. We would like to assess how value-added estimates are affected by measurement error conditionally on school composition, namely, average intake score. We do it graphically and only for the total score of the lower secondary school exam and for the fixed and random effects models. Results for subject specific value-added or different models could be obtained from the author.

On the figure below one can see that the difference between value-added estimates before and after correction is strongly correlated with the average intake score. This means that measurement errors affect school effectiveness estimates in an important and politically non-negligible way. The difference in estimates is trivial only for schools with intake scores near the average. For schools with low achievers correcting intake scores for measurement error increases their value-added while for schools with high achievers the opposite is true. This effect is highly visible for the fixed effects model where strong negative relation is observed. While in the case of the random effects model similar negative relation was estimated there are schools with high-achievers which benefit from the measurement error and which lose because of this kind of error. Similarly, for schools with low-achievers there are both losers and winners. In the case of random effects model the impact of measurement error on ranking is much stronger and is only weakly correlated with school's average intake scores. It should be noted that observed effects are of importance from the practical point of view. The difference between value-added estimates with and without measurement error correction are of great magnitude. In the light of these findings it should be clear that development of proper and reliable testing framework is crucial for any value-added system which is supposed to be used as a mean of assessing and comparing school effectiveness.

Figure 1. Comparison of value-added estimates before and after measurement error correction for the fixed effect and the random effect models.



VIII. How to publish value-added estimates?

Exam results are published at the country and regional aggregate levels. Mean scores are reported to school principals and in some regions to local governments or *Kuratoria* (school inspectorates). No information about statistical significance is given. Schools are located on the stanine scale which gives 1 to 9 ranking for all schools in Poland. Lower secondary school results are reported for both subjects. School principals also receive exam scores for all their students.

We assumed that educational value-added measures should be reported with information about their statistical reliability. It was proposed not to report value-added point estimates but to report confidence intervals. This was a novel idea in the Polish examination system and some were afraid that such information is too difficult to understand for parents and even school principals. However, the expert group working on value-added in Poland pointed out that publishing value-added without confidence intervals is not justified, especially when such measures are used for comparisons (see Goldstein, 1997, for similar view).

Value-added estimates were not officially presented to schools, but first results were discussed with a wide group of stakeholders during the summer workshops in 2006 where tables with predicted scores were used to calculate value-added for hypothetical schools. These tables were published and it is now possible to calculate value-added measures for any school or group of students if merged data from 2002/2005 exams are available. It was explained during the workshops that value-added measures should be reported as interval estimates and a simple method of obtaining confidence intervals was showed. Some basic use of confidence intervals to make statistical interference about value-added differences between schools or freely defined groups of students was explained. At the end of 2006 value-added calculator was made available to the public on the Internet allowing anybody to calculate point estimates with student level data at hand. This was also updated recently for 2004/2007 cohort.

It is not obvious how final implementation of value-added will take into account the problem of statistical significance of estimates. The expert group position is that value-added should be reported as an interval estimate. Obviously this needs further attempts to educate stakeholders about proper interpretation and use of such measures, but employing confidence intervals seems to have at least two important virtues. First, these will limit a possibility to produce league tables which is seen by many practitioners and experts as the potentially most harmful consequence of value-added assessment implementation. Second, these will help to utilize value-added information as a method of self-evaluation for school development reasons and as a tool to assess educational policy programmes on the local or regional levels. Clearly, teaching local actors how to properly use statistical information is not an easy task. However, such efforts are essential when one wants to develop fruitful quantitative assessment system.

IX. Policy relevance of value-added methods

In this section a problem of policy relevance of the value-added definition and interpretation is discussed. In the Polish case, in official reports and during the workshops with stakeholders, value-added of a given school was defined as a quantitative measure of its effectiveness. It was said that such measures could be used to assess teaching quality and how school affects their students knowledge gains. This typical interpretation of value-added measures was challenged by some experts and practitioners. Criticism touched some important points that should not be overlooked. We attempt to summarize them in a few general statements presented below.

a) Value-added models do not satisfactorily measure knowledge growth and at best these are proper methods to estimate exam score gains. This way of criticism is based on the more general idea that external exams measure only part of school teaching efforts. The point

often mentioned is that tests do not measure some important domains which in many cases are non-measureable.

b) Value-added methods do not sufficiently control for school and student level characteristics that affect achievement. Thus, they cannot serve as a basis for school effectiveness assessment. Especially, they should not be used to judge teaching quality if there are important uncontrolled factors affecting knowledge gains.

c) Publishing value-added measures will boost production of rankings which can negatively affect school development and will give further impulses for between-schools student segregation.

The issues listed above clearly touch the problem of utilizing value-added measures in educational policy-making. They need to be addressed if one wants to fully assess the objectives of value-added methodology. The first point (a) is the most general way of criticizing value-added approach to measure school effectiveness based on exam scores. There are two separate problems here. One is a problem of proper definition and measurement of school system outcomes. External exams are built upon centrally set attainment standards which define what knowledge students should achieve in schools. In Poland definition of attainment standards is too simple and still controversial. However, this is not a problem of value-added system alone, but it is general problem of developing satisfactory and widely accepted examination standards. While this standards are established this kind of criticism is against them and not against the value-added system. The second problem is due to a more general critique of assessing student development using exam scores. It is hard to deny a view that it is impossible to produce tests which will measure all possible school efforts or ability gains. We are not going to discuss these issues here, because they are too general and related to educational test measurement rather than value-added estimation. However, it should be noted that examination system in Poland is quite new and still developing. This clearly affects the quality of value-added assessment. This is why publication of value-added measures was postponed and many research projects were conducted to see whether estimates of school effectiveness are related to other characteristics of teaching quality. Nevertheless, the fact is that in such an immature system developing value-added measures is far more difficult and should be done with great caution.

The second point (b) addresses the problem of omitted variables bias. This is a serious problem when one wants to use value-added measures to assess teaching quality. However, its practical importance depends on a way value-added are supposed to be used or on who will use them and for what reasons. If one needs a pure measure of school efforts then any factors that are independent from decisions of school principals or teachers should be controlled for. Obviously, this is never fully possible in observational, non-experimental studies (see Rubin et al., 2004; Goldstein 1997). On the other hand, even in this case possible bias can be negligible and value-added measures could still be used as a reasonable proxies for teaching quality.

Nevertheless, even when controlling for school level inputs and student characteristics is not possible, value-added estimates can be very useful measures of total achievement gains for parents. Parents do not need to know what the pure teaching quality effect on achievement growth is. They just need a measure of overall average gain for students with similar characteristics to their own child. Value-added measures seem to be quite useful in this regard if they are properly reported (see Meyer, 1997; Raudenbush, 2004).

The last point (c) is quite important in a decentralized school system with some elements of free-market competition and school choice. In Poland, school principals have to admit all students from their area but can choose among students from other districts, which can increase between school segregation. In big cities like Warsaw parents often change neighbourhoods to send their child to a better school. On the supply side, introduction of value-added can produce incentives for school principals to concentrate on easy-to-teach students or to concentrate on work that is measured by tests. These problems are of importance when developing value-added methods, but they are even more important in a system when only raw examination scores are available. In this case there is no measure

of teaching quality and parents choose schools according to their composition of students (their average achievement levels). In this case one cannot avoid segregation and cream-skimming. With value-added in place schools could be "good" or "bad" independent from the composition of students. Additionally, there are several ways to limit the extent of unneeded effects. For example, publishing results as interval estimates or even making them non-public will curb the production of rankings. Value-added can be designed to inform about school effectiveness separately for low and high-achievers (e.g. by employing random slope models). These attempts will limit incentives to concentrate solely on better or easier-to-teach students.

X. Value-added analysis as a policy evaluation tool

Value-added models could be employed to assess policy effects or evaluate teaching effectiveness for freely defined groups of students or schools. While the simplest approach in this case is to compare averages of value-added scores among groups of interest, the more sophisticated analysis could use value-added regression models where many kinds of analysis are possible. Using the value-added model for policy evaluation could be a very fruitful exercise. Basically, collecting the data for the value-added analysis is usually the most demanding task. If the data are in place and value-added analysis for a school assessment system was already done, then it is straightforward to use this data for policy analysis. Policy makers should be aware of this possibility when thinking about implementing value-added systems. While value-added models do not provide causal estimates and one needs to conduct experimental or quasi-experimental research to obtain clear evidence on causal links, the value-added approach is much more reliable than usual cross-sectional analysis where the levels of achievement, and not the growth, are used to assess policy effects. Taking into account intake scores of students limits the bias caused by e.g. correlation of compositional effects with resources or policy programmes. Value-added models provide also directly interpretable results when teaching quality is to be assessed which could be of importance for policy makers.

We do not want to discuss in details how value-added models could be used in policy evaluation but would like to present a simple example of such analysis which addressed important policy question. Namely, we discuss a research where the main empirical question was whether decentralized educational expenditure affects teaching quality. The goal of the research was to test if spending per student in lower secondary school affects achievement growth. Below, we provide some evidence from value-added modelling that decentralized expenditures do not affect teaching quality. This finding is in line with many papers in the economics of education literature which suggest that expenditures, or other inputs, have no impact on teaching outcomes, but there are still methodological controversies which we are not going to address here (see opposite views in: Hanushek, 2003, and Krueger, 2003). Note, however, that this research is devoted to the analysis of decentralization where value-added modelling is a novelty and produces interesting evidence with not only quantity but also quality of services considered.

For detailed description of decentralization in Poland and data used in this analysis readers are referred to other papers (see Jakubowski, Topińska, 2007; Jakubowski, 2007b). It is enough to mention that the dependent variable in the model is a score obtained at the end of lower secondary school and as an intake score we used test results at the end of primary school. Expenditures were calculated as the average spending per student during the period of education in lower secondary school (deflated using HICP Eurostat index). We used three-level model with random effects at the local government and at the school level. Additionally, we used data for three cohorts which finished lower secondary school in 2005, 2006 and 2007. This way within- and between-variance was explored making results more robust to hidden characteristics of local governments and schools. The estimated random intercept value-added model is given by the equation below:

$$y_{isg} = \pi_0 + \mathbf{C_{isg}}\boldsymbol{\beta}_\mathbf{k} + \pi_1 x_{isg} + \eta exp_g + u_g + v_{sg} + \varepsilon_{isg} \quad (5)$$

where y_{isg}, is the score of i-th student on the exam at the end of primary or lower secondary school s in the gmina g, x_{isg} is the intake score of i-th student, $\mathbf{C_{isg}}$ is a vector of individual, school and gmina explanatory variables, exp_g is an 3-year-average expenditure per lower secondary school student in gmina g, ε_{isg} is the individual error term, u_g is the gmina random effect and v_{sg} is a school random effect. As usual it assumed that $\varepsilon_{isg} \sim N(0,\sigma_\varepsilon^2)$, $u_g \sim N(0,\sigma_u^2)$ and $v_{sg} \sim N(0,\sigma_v^2)$. The parameter of interest is η which measures the effect of expenditures on student achievement growth in lower secondary education. The positive estimate of this parameter means that higher expenditures are positively affecting teaching quality. In all regressions we added also year dummies and interaction terms between these dummies and intake scores to allow differences in slope between years. We also added quadratic term of intake scores to fit the non-linear relation.

Results are presented in the table below. All explanatory variables are described in the first column which clarifies the specification of the regression. As we said already, value-added models incorporate intake scores and could be used to estimate impact of any factor of interest on achievement growth, not the level. To see the difference consider column (1) and (2) in the table below. In the first column a simple model without intake scores and with few individual variables, random effects for gmina and lower secondary school is presented. We excluded from the table estimates for being dyslectic or winning the "science Olympics", because these were not interesting, however, they were present in all regressions. The second column contains estimates for a simple value-added model. The only difference is the inclusion of intake scores (with quadratic term and interacted with year dummies to limit the impact of year-to-year changes in score distribution). Note, that unexplained variance is much lower in the second model at all levels: gmina, school and individual. Inclusion of intake scores visibly improve fit of the model.

Compare now columns (3) – (7) were expenditures per student in lower secondary school were included and additional controls were added in each column. Note, that estimate of the effect of expenditures is negative and almost the same in all regressions. Thus, one can conclude that increasing school expenditures decrease teaching quality. While the negative sign could be still caused by endogeneity problems, we want to emphasize that from the practical point of view these effect is simply negligible or non-distinguishable from zero. Estimate suggest that each additional 1000 PLN (around 400 USD, note that country average is less than 4000 PLN) decreases student scores by around 0.07. With standard deviation of individual test scores around 17 and standard deviation of gmina's average test scores around 3.8 this effect should not be considered as important for policy makers. In fact, it shows that differences in decentralized expenditures have no visible impact on teaching quality.

Other variables are difficult to interpret, because they serve more as controls rather than factors which are expected to directly affect achievement growth. Note, however, that non-public schools show significantly higher quality despite the control for intake scores. This seems interesting and should be analyzed more carefully in future research to find why these schools achieve more than public ones. Is it the effect of additional private resources, different organization, methods of teaching, parental involvement, or higher teacher salaries? All these factors are similarly probable and we cannot separate them in this research.

Table 5. Value-added lower secondary school regression results.

Dependent variable: **Total score from the lower secondary school exam**							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Gender	2.823***	0.164***	0.165***	0.164***	0.164***	0.156***	0.157***
	(0.030)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
Intake score		**1.254*** **	**1.256*** **	**1.252*** **	**1.252*** **	**1.253*** **	**1.253*** **
		(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Intake score * year = 2006		-0.012***	-0.012***	-0.012***	-0.012***	-0.012***	-0.013***
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Intake score * year = 2007		0.042***	0.042***	0.042***	0.042***	0.043***	0.043***
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Intake score ^2		0.012***	0.012***	0.012***	0.012***	0.012***	0.012***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Lower secondary school expenditure per student			**-0.067*** **	**-0.057** **	**-0.076*** **	**-0.072*** **	**-0.066*** **
			(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
Non-public school			2.569***	3.204***	3.179***	2.940***	2.934***
			(0.167)	(0.174)	(0.178)	(0.220)	(0.220)
Natural log of school size				0.532***	0.544***	-0.170**	-0.173**
				(0.042)	(0.044)	(0.058)	(0.058)
IQR of intake scores				-0.035***	-0.035***	-0.045***	-0.045***
				(0.007)	(0.007)	(0.007)	(0.007)
Natural log of a number of students in gmina					-0.046	0.372***	0.259***
					(0.047)	(0.058)	(0.059)
Natural log of gmina income per citizen					1.422***	1.388***	1.555***
					(0.209)	(0.212)	(0.213)
Primary school expenditure per student							-0.337***
							(0.035)
Preschool participation rate							0.008***
							(0.002)
type of area dummies						Yes	Yes
region dummies			Yes	Yes	Yes	Yes	Yes
year dummies	No	Yes	Yes	Yes	Yes	Yes	Yes
Constant	54.509***	15.354***	14.941***	13.090***	2.719	3.927*	4.268**
	(0.105)	(0.107)	(0.188)	(0.267)	(1.549)	(1.592)	(1.588)
Random part estimates							
SD of gmina intercept	2.12	1.22	0.90	0.95	0.96	0.94	0.95
SD of school intercept	6.53	3.03	2.90	2.86	2.85	2.85	2.84
SD of residuals	16.86	10.34	10.34	10.34	10.34	10.34	10.34
Log restricted-likelihood	-5633162	-4983968	-4979920	-4979837	-4979816	-4979806	-4972839
N of students	1325059	1325059	1324076	1324076	1324076	1324076	1322255
N of schools	6212	6212	6211	6211	6211	6211	6211
N of gmina	2467	2467	2466	2466	2466	2466	2466

Note: *** denotes significance at 1%, ** at 5%, and * at 10% level.

Negative impact of interquartile range of intake scores is what was expected and suggests that schools with more heterogeneous achievement levels of students have lower teaching quality. It is not our goal to discuss all potentially valid explanations of why this is the case in Poland or in general, but it seems to be a fruitful exercise to research this effect in details in future. School and gmina student population size effects are clearly correlated with other characteristics of gmina which affects outcomes, mainly with the size of gmina population. Inclusion of type of area dummies (last two columns) switched the sign of these effects showing that these are proxies for gmina size. Similarly, logarithm of gmina income should not be interpreted as the effect of additional gmina resources, because those were already included, but more as a proxy for citizens' wealth and other characteristics correlated with it (e.g. parental education or professions). Finally, primary school inputs and preschool participation were included in the regression. While preschool participation effect was found to be positive as it was expected, the negative sign of expenditures contradicts earlier findings. It is probably correlated with area size or other characteristics and should not be directly interpreted for reasons already mentioned.

We also estimated several random slope models where the impact of chosen factors on the slope of intake scores in a gmina was of main interest. We tested the hypothesis that higher expenditures increase a chance of low achievers to learn similar amount of knowledge as their high achieving peers. More specifically, we were interested in the interaction between intake score slope and expenditures. Thus, we estimated random slope model where intake score slopes were allowed to vary between gmina and with interaction term explaining slopes by the level of expenditures. Positive estimate of this interaction term suggests that expenditures have equalizing impact. The regression is given by equation below:

$$y_{isg} = \pi_0 + \mathbf{C_{isg}}\boldsymbol{\beta}_{\mathbf{k}} + \eta_0 exp_g + (\pi_1 + \eta_1 exp_g + \varsigma_g)\tilde{x}_{isg} + u_g + \varepsilon_{isg} \qquad (6)$$

where $\tilde{spr}_{isg}$ is an intake score centred around the „grand mean", π_1 is an average intake score slope, η_1 is a parameter of interest which reflects the effect of expenditures on intake score slope, and ς_g is a random effect which allows slope to vary between gmina. Negative sign of the η_1 is expected if the expenditures have "equalizing" effect on learning outcomes. For computational reasons we estimated this model only with gmina level random effects assuming as usual that random intercepts and slopes are normally distributed and that $\mathrm{cov}(u_g, \varsigma_g) \neq 0$ which means that we didn't specify the correlation between slopes and intercepts allowing them to freely vary.

Note, that in this case intake scores have to be centred around some value to obtain meaningful results. Centring around the "grand-mean" (population mean) produces the same model but with estimates directly interpretable as effects for the average student. However, centring around some other location gives different model (de Leeuw, 2005). In our case we centred intake scores around the grand-mean but also around the 10th percentile of intake score calculated separately for each gmina. Thus, in the latter case we estimated the impact of expenditures on low achieving students defined relatively to the population of students in each gmina. Additionally, we added interaction term between expenditures and intake scores explaining random variation of intake score slopes at the gmina level.

The results are presented in the table 3 below. Columns (1) and (3) contain two models with intake scores centred around grand-mean. Columns (2) and (4) contain similar models but with intake scores centred around the 10th percentile. For clarity, we omitted from the table several control variables used in estimation, namely, all individual characteristics present in earlier value-added models except intake scores, dummy for non-public schools, and indicators for regions. In addition to earlier models we added school's average intake score to regressors to better fit the nonlinear relation between intake and lower secondary score and to take into account so called compositional effects (Goldstein, 1997)

First note, that estimated impact of expenditures on random school intercepts is negative despite the way intake scores were centred. This means that expenditures similarly affect average students and low-achievers. As we noted already these effects are non-distinguishable from zero from the practical point of view. Additionally, we were also interested in the estimate of the interaction term between expenditures and centred intake scores. Our estimates of this interaction term were consistently negative which means that increasing expenditures lowers intake score slope. This can be interpreted as equalizational effect of expenditures because in lower secondary schools on which local governments spend more low achievers progress more. However, these effects are very weak and from the practical point of view are negligible. Thus, the findings could be summarized as showing no effect of gmina expenditure on the teaching quality. This seems to be true despite the level of students. Huge sample and proper statistical models produced very accurate estimates which makes these findings precise and robust.

Table 6. Value-added random slope lower secondary school regression results.

Dependent variable:				
Total score from the lower secondary school exam				
	(1)	(2)	(3)	(4)
Intake score grand-mean centred: columns (1) (3) 10th percentile centred: columns (2) (4)	1.254*** (0.010)	1.240*** (0.010)	1.239*** (0.010)	1.213*** (0.010)
Intake score * year = 2006	-0.009** (0.003)	-0.009** (0.003)	-0.009** (0.003)	-0.009** (0.003)
Intake score * year = 2007	0.050*** (0.003)	0.049*** (0.003)	0.049*** (0.003)	0.047*** (0.003)
Intake score^2	0.013*** (0.000)	0.013*** (0.000)	0.013*** (0.000)	0.013*** (0.000)
Lower secondary school expenditure per student	**-0.074***** (0.020)	**-0.055**** (0.024)	**-0.097***** (0.020)	**-0.105***** (0.024)
Expenditures * intake score	**-0.005**** (0.002)	**-0.001** (0.002)	**-0.004*** (0.002)	**-0.002*** (0.002)
Natural log of school size			0.279*** (0.022)	0.279*** (0.022)
Natural log of a number of students in gmina			-0.207*** (0.054)	-0.172*** (0.055)
Natural log of gmina income per citizen			1.589*** (0.220)	1.648*** (0.221)
IQR of intake stores			-0.128*** (0.006)	-0.128*** (0.006)
School's average intake score			0.179*** (0.006)	0.180*** (0.006)
Constant	14.553*** (0.258)	4.372** (1.654)	-0.201 (1.67)	-0.215 (1.67)
Random part estimates				
SD of intake score slope (gmina level)	0.11	0.10	0.11	0.11
SD of gmina intercept	2.38	2.32	2.41	2.36
COV(slope, intercept)	0.31	-0.17	0.30	-0.18

SD of residuals	10.47	10.47	10.46	10.46
Log restricted-likelihood	-4992993	-4993013	-4991941	-4991958
N of students	1324076	1324076	1324076	1324076

Note: Some regressors were not presented in the table. *** denotes significance at 1%, ** at 5%, and * at 10% level.

XI. Summary

In this report policy implementation of value-added models of school assessment in Poland was presented. We discussed the policy objectives, proposed value-added methodology for lower secondary schools, impact of statistical modelling and measurement error on school effectiveness estimates, and finally we described application of value-added models in policy evaluation. Emphasis of the paper was on the relation between statistical problems and ease or usefulness of implementation. It is now clear that simple and more complicated models produce highly comparable estimates and from a practical point of view simpler models could be preferred for policy reasons. However, there is still room for improvement of validity and reliability of value-added statistical methods and we mentioned several problems that could be solved with better data or more advance models.

Value-added assessment of schools is highly valuable when examination results are already used to compare schools. In this case, the value-added methodology produces more adequate information about teaching quality. Despite some methodological problems discussed in the paper, value-added models are much more reliable and not that misleading as raw examination scores used as a mean for ranking-based comparisons. The paper shows, however, that not only details of statistical procedures are of importance here. In practice a way the value-added estimates are published and used locally is much more important than any of statistical problems mentioned. We pointed out that publishing point estimates is not that helpful as providing general public with properly presented interval estimates. Schools can benefit a lot if value-added system allows them to conduct their own analysis based on value-added methodology and data. The value-added statistical exercise is not a purely scientific undertaking. It should be conducted with a clear idea of how it could be used and its value depends on how it interacts with its final recipients: parents, teachers, and policy-makers.

References

Baltagi B. H. (2005). Econometric analysis of panel data. 3rd ed. New York: Wiley.

Dolata R. (2006). Informacja o pracy zespołu metodologicznego ds. prognozowania i komunikowania wyników egzaminów zewnętrznych. Research Bulletin „Egzamin" 8/2006, Central Examination Board (www.cke.edu.pl)

Goldstein H. (1997). Methods in school effectiveness research. School effectiveness and school improvement. 8: 369-95.

Goldstein H., Huiqi P., Rath T., Hill N. (2000). The use of value added information in judging school performance. British Educational Research Journal, Vol. 28, No. 4 (Aug., 2002), pp. 632-633

Jakubowski M. (2006a). Metody szacowania edukacyjnej wartości dodanej. Research Bulletin „Egzamin" 8/2006, Central Examination Board (www.cke.edu.pl)

Jakubowski M. (2006b). Empiryczna analiza metod szacowania edukacyjnej wartości dodanej dla gimnazjów. Research Bulletin „Egzamin" 8/2006, Central Examination Board (www.cke.edu.pl).

Jakubowski M. (2007a), Wpływ czynników ekonomicznych na wyniki egzaminów zewnętrznych, Research Bulletin „Egzamin" 12/2007, Central Examination Board.

Jakubowski M. (2007b), Decentralization and teaching quality, unpublished paper.

Jakubowski M. (2007c), Volatility of value-added estimates of school effectiveness, forthcoming.

Jakubowski M., Topińska I., 2007, Impact of Decentralization on Public Service Delivery and Equity. Education and Health Sectors in Poland 1998 – 2003, UNDP research report, CASE – Center for Social and Economic Research.

Ladd H., Walsh R. (2002). Implementing value-added measures of school effectiveness: getting the incentives right. Economics of Education Review 21 (2002) 1–17

McCaffrey D., Lockwood J., Koretz M., Hamilton L. (2005). Evaluating Value-Added Models for Teacher Accountability. Rand Corporation MG-158.

Meyer R. (1997). Value-Added Indicators of School Performance: A Primer. Economics of Education Review, Vol. 16, No.3, s. 283-301.

O'Brien P., Paczynski W. (2006). Poland's Education and Training: Boosting and Adapting Human Capital. OECD Economics Department Working Papers 495, OECD Economics Department.

Rabe-Hesketh S., Skrondal A. (2005). Multilevel and Longitudinal Modeling Using Stata. Stata Press.

Raudenbush S. (2004). What are value-added models estimating and what does this seem to imply for statistical practice? Journal of Educational and Behavioral Research, 29, 121–130.

Rubin D., Stuart E., Zanutto E. (2004). A potential outcomes view of value-added assessment in education. Journal of Educational and Behavioral Research, 29, 103–116.

Stata (2007). Longitudinal/Panel-Data Reference Manual. Release 10. Stata Press.

Maciej Jakubowski
Faculty of Economic Sciences, Warsaw University
Dluga 44/50, 00-241 Warszawa, Poland
email: mjakubowski@uw.edu.pl