BULGARIA

Piloting Statistical Models for Estimation of Schools’ Value-Added Using the Results from the National Assessments

*Key results and findings*

March 2015

WORLD BANK
EUROPE AND CENTRAL ASIA REGION
EDUCATION
Acknowledgements

This report was prepared under the Programmatic Education Sector Technical Assistance Program of the World Bank (WB) in response to the request of the Bulgarian Ministry of Education and Science (MES) to pilot and document the implementation of a set of statistical models for estimation of the value-added of Bulgarian schools using data from the national census-based student assessments. The pilot and the report were prepared by a team led by Plamen Danchev (Education Specialist, World Bank) and comprising the following researchers from the Bulgarian Association for Educational Measurement and Evaluation (BAEME): prof. Kiril Bankov (Lead Consultant, BAEME), Vessela Stoimenova (Consultant, statistical analyses, BAEME), Dimitar Atanasov (Consultant, statistical analyses, BAEME).

The scope of the pilot implementation, the key parameters for the activity and the required data for the analysis were jointly defined by the WB and a working group of MES experts led by Mr. Krassimir Valchev, Secretary General of the MES and comprising Evgenia Kostadinova (Director for Curriculum and School Programs, MES), Orlin Kouzov (Director for IT in Education, MES), Neda Kristanova (Director, CKOKUO – National Testing Agency), Penka Ivanova (Director of Policy Directorate, MES), Sashko Arabadjiev (Expert, CKOKUO), Rumiana Tomova (Expert, MES), Stela Mitsova (Expert, MES) and Radosveta Drakeva (Adminsoft, developer of the Education Management Information System of the MES).
Summary

At the request of the Ministry of Education and Science (MES), the World Bank (WB) has expanded the scope of the schools value added measurement pilot using test results from the national census based student assessments. The first pilot was implemented in 2013. The present report summarizes the results of the second phase of the pilot.

The value added measures (VAM) of school performance try to isolate the contribution of the schools to the academic growth of the students from the factors that are beyond the control of schools teachers and management. As such, the VAMs are arguably among the most precise and equitable test-based measures of school performance, as long as their estimation is based on valid and reliable student assessments and appropriate statistical models.

The results from the present pilot show that the three tested models (a fixed effect linear regression model, a random effects model and a random effects model benefitting from the inclusion of the language spoken at home by the observed students) could be used for the purposes of estimating the value added of Bulgarian schools, but the model with the most advantages is the random effect model using the “language spoken at home” variable. It produces statistically more precise composite VAMs based on the two subjects (Math and Bulgarian language and literature) and allows a problem-free adjustment of small schools VAMs.

To counteract the negative impact of the small number of tested students in small Bulgarian schools on the estimation of their value added, a statistical shrinkage of the VAMs is recommended for all schools with less than 20 tested students per cohort. The two random effect models allow a precise way of implementing such shrinkage.

The successful introduction of VAMs in Bulgaria would benefit from the following changes to the existing national assessment framework:

- Enforce the recently announced video surveillance not only for the Matura exams, but for all national assessment tests;
- Increase the number of items to achieve better reliability of the assessment tests. This is especially important for the subjects that are used for the VAM analysis, namely Bulgarian language and literature and Mathematics.
- Include items measuring both lower- and higher-order competencies embedded in the national curriculum, not just the presently tested achievement of the minimum standards. The “minimum standards” approach creates a “ceiling” effect for the students with better performance due to the low discriminating power of the tests and affects negatively the estimation of the value added.
- Related, the scales of the tests should be revisited to ensure there is a sufficient stretch to measure the progress of both very low and very high achieving students;
- Consider making the Matura exam in Math compulsory so that high schools are also included in the VAM framework and a composite VAM could be produced by using the Matura results from both Math and Bulgarian Language and literature.
Introduction

1) The objective of this report is to document the process and summarize the results from the second phase of the pilot implementation of statistical models for measuring the value-added of Bulgarian schools through analysis of the national student assessments results. The first effort to conduct such analysis was made in 2013 as part of the World Bank’s technical assistance program for the education sector in Bulgaria. The first phase of the pilot suffered from a number of significant constraints, such as the large share of missing data for the test results of the students, the failure to collect contextual data for the entire analyzed cohort of students, and the resulting limitations for quality value added analysis. At the request of the Ministry of Education and Science (MES), the work has been expanded in 2014 and 2015 to pilot new statistical models with a full set of data for two consecutive student cohorts. This report presents the technical aspects of the pilot and the key outcomes in terms of value-added measure for each of the schools included in the analysis. It is intended to ensure the institutional memory for the pilots and to stimulate the discussions at technical level among experts and decision makers at the MES, focusing on the relevance and applicability of the piloted statistical models, in the context of Bulgaria student assessment framework, and given the available data. The report documents as well all the data processing, adjustments and procedures run as part of the school value added modeling. In this respect, it is also intended for expert statisticians and researchers.

2) Test based measures of school performance have been used for long time in many education systems, mainly in the form of average scores on standardized tests. There has been growing recognition of the limitations of this approach and the unintended consequences it could bring about if raw unadjusted absolute scores are used for accountability or are attached to “sticks and carrots” type of incentives for school performance. These measures often do not take into account other factors that influence educational achievement, such as: the starting point and native abilities of students; their socio-economic background; the influence of peers and individuals in and outside school; various events and situations that occur outside the school that might affect student learning; etc. One option is to use a simple student growth percentile model for assessing the contribution of the schools to students’ outcomes. This approach, however, has significant limitations compared to the VAMs.

3) International experience suggests that the VAMs based on a valid and reliable student assessment system are the most equitable and precise test-based indicators for school performance developed to date. The school’s value added represents a measure of school performance derived from statistical analysis of the raw test scores of one and the same cohort of students in at least two points in time, further adjusted to control for the socio-economic characteristics of the students. Through such analyses the students’ achievement could be decomposed into components attributable to schools and components attributable to the students’ background characteristics. The discrete or more direct use VAMs for accountability or school improvement purposes is gaining popularity and policymakers increasingly accept them as a measure of school performance (MILO, 2008). Among the EU member states, well-established VAMs exist in England (Ray, 2006) and Poland (Jakubovski, 2008). In both countries VAMs are used for school improvement, as well as for accountability purposes.

4) The goal of the MES to develop fair and objective measures for school performance that could be publicly disclosed has stimulated discussions among stakeholders on the possibility to develop value-added assessment models based on analysis of student assessment data as part of a broader set of indicators measuring the quality of schools.

Key prerequisites for the introduction of the school value-added measures in Bulgaria

5) External standardized national assessment tests were first introduced in Bulgaria in 2007. Presently national assessments are census based (covering all students in the tested cohorts), administered annually at the end of elementary (Grade 4), primary (Grade 7) and upper secondary (Grade 12, Matura) levels. The new draft law on Preschool and School Education envisions a new lower secondary education stage covering Grades 7-10, and a national, census-based, annual assessment at the end of Grade 10. Presently, the national student assessment results in Bulgaria are reported on a non-transformed raw point scale accompanied by a table that allows a transformation of the raw points to the six-grade rating scale, adopted

---

1 OECD, Measuring Improvement of Learning Outcomes
for measuring students’ performance. Based on this, each school receives an average score and some means of comparison with other schools in the country. However, datasets with the results of all schools in Bulgaria are not publicly available to avoid improper comparisons. The assessment results have not yet been put to a use that can significantly improve decision-making, policy development and incentivizing school performance gains. The census-based assessment in Bulgaria needs to provide education stakeholders with information about the performance progress of individual schools and to facilitate the implementation of national policies for improving the quality of education. VAMs could serve well the above objectives.

6) Linger ing concerns remain, however, about the supervision of the test taking process and the arrangements to counteract gaming in test taking. The Government has recently announced measures to improve the integrity of the Matura tests by using video surveillance during test taking and taping the process for subsequent scrutiny. This measure need to be enforced and implementation should cover all national assessment tests beyond the Matura. Until then, the returns to Government investment in the assessment system would be lower than its potential. Any attempt to use the assessment results and the related VAMs for accountability or school improvement purposes will be inappropriate, given the integrity concerns. While analyzing the gaming in tests is beyond the scope of the present report, it should be stressed that rolling out the VAMs in Bulgaria without resolving these issues will result in unfair distribution of recognition and resources across the system. The Bulgarian Government needs to first address the above mentioned shortcomings, ensuring that test results do reflect in objective manner the knowledge of tested students and that the test instruments are sufficiently improved to fully respond to the requirements of VAMs.

7) Further, the design of the test instruments, need to be revisited to include items capturing both lower- and higher-order competencies of the national curriculum, not just the minimum standards, which presently form the ceiling of competencies measured by the 4 grade tests. Related, the scales of the tests should be developed according to the commonly accepted rules and should have sufficient stretch to measure the progress of both very low and very high achieving students in order to produce good value added analysis (Sanders 2003). Otherwise, “floor” or “ceiling” effects may lead to underestimating the value-added of high performing schools or not adequately measuring developments in low-performing schools (Ray 2006). As shown in the graphs and histograms in the next chapter, the test results after grade 4 are far from the normal Gaussian distribution. This is a clear sign that the tests do not have sufficient discriminating power for the high achieving students. The reliability of the assessment tests can be improved by including more items. This is especially important for the subjects that are used for the VAM analysis, namely Bulgarian language and literature and Mathematics. However, including more items may have budgetary implications. If changes to the national assessment instruments are considered and more items are added, the testing time will increase accordingly. Further, the production of more items in a bid to avoid the so called “floor” and “ceiling” effects will also increase the budget of the national assessments.

8) As part of the dialog with the MES on the present VAM pilot, a number of features and parameters were discussed and agreed upon, including the following:

a) Producing separate VAMs for the different educational stages. Value-added estimates are produced for the schooling confined between the grade of the prior attainment measure and the grade of the outcome measure in a particular subject or group of subjects. As such, the value-added modeling used in Bulgaria will produce value added measures in Bulgarian language for each education stage – in grade 7, marking the end of primary education (based on prior attainment in grade 4), in grade 10, marking the end of lower-secondary education (based on prior attainments in grades 4 and 7, or grade 7 only) and grade 12 (based on prior attainments in grade 4, 7 and 10, or just grade 10). Similar value-added measures could be produced for cognitive attainment in mathematics. Further, to arrive at school value-added estimates rather than subject value-added estimates, results for the two tested subjects will be aggregated and could therefore be presented as a composite value of the school value added for each educational level. It should be noted that for this purpose test results in both subject – Bulgarian language and literature and Mathematics – are needed at the end of Grade 12 in addition to the test results after Grade 7. At present, these are available only for Bulgarian language and literature, since the Matura is not compulsory for Mathematics. Should VAM be expanded to include upper secondary educational level, MES should consider the introduction of compulsory Matura in Mathematics. For the purposes of the present VAM pilot, value added is produced only for primary education (comparing assessment results from grade 4 and grade 7). This means that basic schools
(grades 1 to 7) may be included in a value added calculation once, for their value added estimated for the 4-7 grade span; the general comprehensive schools may have 3 value added measures (first, for the 4-7 grade span; second, for the 7-10 grade span, and third, for the 10-12 grade span) allowing them to participate in value added school rankings for the three key stages of education. In order not exclude schools offering 1-4 grades from the overall accountability system, results from the tests in grade 4 may be used in a status contextualized attainment model (provided adequate contextual student data is available) and adjusted results used according to the policy objectives set for elementary education.

b) **Vertical and horizontal linkage of testing instruments.** There is considerable debate about the comparability of test scores and the conversion of scores into meaningful and comparable scales (Braun 2000; Dorans et al. 2007; Patz 2007; Kolen and Brennan 2004). Longitudinal value-added analyses typically employ test score scales that have been vertically linked across grades (Harris et al. 2004). Different strategies to carry out vertical linking yield score scales with different properties that, in turn, can have a substantial impact on value-added estimates (Patz 2007). The construction of tests that allow for vertical calibration and establishment of clear relationships between what is tested in the different grades and its expression on common scales suggests that the ideal setting for value-added modeling would be assessments tests based on the Item Response Theory. The latter is best poised for vertical linking and generally produces tests of higher reliability.

In practice, however, many value-added models (such as the ones implemented in England and Poland) do not require that the test scores be vertically scaled. They simply require that scores in successive grades or educational stages tested be approximately linearly related and, in most cases, that is a reasonable measure (Doran and Cohen 2005). In England (Ray 2006; Evans 2008) and Poland (Jakubowski 2008) the tests are based on the classical test theory and are aligned with curricular objectives and their scoring with the prescribed levels of proficiency, which allows for optimal implementation of VAM. Placing the national assessment tests in Bulgaria on the same scale and achievement continuum, and the extent to which the Matura exams are, or can be, placed on the same scale and continuum, are important issues that need to be resolved in a manner that allows quality implementation of value added analyses.

c) **Value-added of small schools.** The accuracy of VAMs may suffer if estimates are based on a limited number of student assessment data. While this may not be an issue for most schools, small primary or basic schools’ VAMs may vary significantly from year to year because of the small number of students enrolled due to the substantial sensitivity of the measure to changes in cohorts’ composition. This issue is examined at length in Ray (2006) using data for small schools in England. One solution that was discussed and agreed with the MES is to perform shrinkage of test results (as part of the specifications of the value-added statistical model), thus adjusting small schools’ value added estimates towards the mean. The models presented in this report have analyzed VAM based on two definitions for small schools – one, a school is considered small when there are less than 30 students per a tested cohort, and two, less than 20 students per a tested cohort. Shrinkage was performed for all these schools.

d) **Technical complexity of value-added modeling.** Value-added models range from rather simple regression models to extremely sophisticated models that require rich data sets and state-of-the-art computational procedures. In general, it could be argued that more complex models do a better job of yielding estimates of school performance that are free of the influence of confounding factors, although there is still some argument on this point. The disadvantage is that, typically, the greater the level of complexity, the greater are the staffing requirements and the longer is the time required to set up and validate the system. More complex models usually require more comprehensive data (years and subjects), so that data availability limits the complexity of the models that can be considered. In addition, the greater difficulties of communicating the workings and use of more complex models might reduce the transparency of the system and increase the problems of gaining the support of stakeholders. As such, it was agreed with the MES to use simpler statistical models, namely a linear regression model and a random effect model.

e) **Low impact on parental choices.** Even though VAM’s are equitable measures and describe better the effort of teachers to improve students’ performance, international research shows that parents are far more interested in the absolute achievement levels for a given school, rather than its value-added. As
such, the impact of VAMs on parents’ choice is likely to be low. The experience in Chile, analyzed with several specifications by Mizala and Urquiola (2013), concludes that parents react to rankings for school choice rather than measurements of value-added in schools.

f) **Unintended consequences that could be triggered by the use of VAM.** As discussed earlier, attempts to cheat during test taking in a bid to improve students’ scores has been the most common form of manipulation of tests results in Bulgaria even without attaching particular incentives to the test results. Such manipulation is also the easiest one to counteract through strong test taking arrangements and careful selection of test supervisors. However, there are a number of documented international instances where various school indicators and high-stakes tests can and have been manipulated in a more sophisticated manner that creates sub-optimal outcomes, including when value-added measures are used to trigger various forms of performance rewards (Nichols & Berliner 2005). When two assessments are employed to estimate schools’ value-added, e.g., grade 4 and grade 7, a school’s value-added increases if there is a larger positive difference between the two assessments; hence there is an incentive both to lift students’ scores in grade 7 and to lower the scores (of those same students) in grade 4. This could be achieved by advising students not to take the grade 4 assessment as seriously as might otherwise be the case or even by encouraging them to deliberately under-perform. More radical actions could include structuring the curriculum so students are not properly prepared for the grade 4 assessments. Yet, strategies can be developed to reduce the likelihood of such sub-optimal activities. For example, the perverse incentive effect could be countered by imposing performance targets for the grade 4 assessment. More generally, schools should have an incentive to lift the performance of students in all assessments, thereby aligning their interests with those of the students. This can be achieved most simply when each assessment is both a prior and final assessment (e.g., grade 7 assessment is an outcome assessment for value added measure for grades 4-7 schooling and a prior attainment for the value added measure for grades 7-10 schooling).

**Results from the pilot**

**Data used for the pilot implementation of the school value-added models**

9) Conducting a value-added assessment requires test scores from at least two separate points in time. As described above, in Bulgaria there are three census-based national assessments. For this analysis tests scores and student level data are used for the cohort of students that took both the national assessment test in Grade 4 (back in 2010 and 2011) and the national assessment tests in Grade 7 (respectively in 2013 and 2014). This makes it possible to use the results of one and the same students of two consecutive cohorts at these two points in time in order to pilot value-added measure of the schools between grades 4 and 7. The national assessments in Grades 4 and 7 cover a number of subjects. The stable and common subject base, however, is observed in Bulgarian language and literature and Mathematics, therefore only the test scores from these two subjects are included in the value-added analysis.

10) The students in the merged dataset are linked by their national ID number (called EGN). This number contains the information about the date of birth. In addition to the test scores of the 4 graders of the 2010 cohort, the Grade 4 national assessment data also contains information about the language spoken at home. The models are piloted using the available data for each of the two cohorts: tests scores from the national assessments of one and the same cohort of students and a limited set of student and school level characteristics (age, gender, type of school, language spoken at home).

11) The data provided by MES are structured in three files. The first file contains information about the results from exams and additional information about student’s language spoken at home. The data are presented in table structured in the following way:

- exam_year – the year of the exam,
- exam_class – grade of the exam,
- school – school where the exam is given,
- student_id – code of the student,
- examcode – code of the exam,
- totaly1 – result_1 of the exam,
- totaly2 – result_2 of the exam.
• totalpoints – result_3 of the exam,
• languageid – code of the language spoken at home.

Thus for each student in the survey there are 4 rows in the data file.

12) The second file contains information about the student’s characteristics as gender, birth year, etc. and additional information about the mobility of the student from one school to another.

• exam_year – the year of the exam,
• exam_class – grade of the exam,
• school – school where the exam is given,
• student_id – code of the student,
• gender – student’s gender,
• birth year – student’s year of birth,
• interim_school – school where the student was enrolled in grades 5/6,
• interim_year – year when the students was in grade 5/6,
• interim_class – grade of the student, 5 or 6,
• EdForm – form of the education of the student in grade 5/6,

13) The third file contains information about the schools.

• School_ID – coded of the school,
• School_Name – name of the school,
• Obl_ID – region where the school is situated,
• Municipality_ID – municipality where the school is situated,
• EKATTE – code of the village where the school is situated,
• cat – category of the village where the school is situated,
• School_Type_1 – type of the school (education),
• School_Type_2 – type of the school (ownership),
• BudgetFrom – type of funding of the school,
• IsProtected – indication for protected school,
• Shifts – number of shifts of instruction,
• NumOfPedag – number of the teachers in the school.

Data processing

14) As a first step the information from the different sources needs to be transferred in one data set. This is done by a program on Perl 5 programming language developed by the authors for the purpose of the study. The data is imported in two hash tables, one for the students, with key student ID, and another for the schools, with key school ID. The language spoken at home has been numerically coded as follows: Bulgarian (Code 1), Roma (Code 2), Turkish (Code 3), and other (Code 4).

15) The obtained data structures are saved as a text file with “comma separated value” format with the following structure:

1. 7-th class score,
2. 4-th class score,
3. Gender of the student,
4. School ID,
5. Language spoken at home,

Descriptive statistics

Student Level Characteristics

16) The important characteristics at student level are: language, spoken at home; gender; mathematics and Bulgarian language score from the national assessment at the end of Grade 4; mathematics and Bulgarian language score from the national assessment at the end of Grade 7.
17) The proportions of the students for different values of the language (language code 1, 2, 3, 4) are presented on Figure 1. For nearly 80% of the population the language spoken at home is Bulgarian.

![Figure 1. Distribution of the language spoken at home](image)

18) The distribution of the percentages of the students’ gender (0-girl; 1-boy) is presented on Figure 2.

![Figure 2. Distribution of the gender](image)

**The 4-th grade level**

19) At the 4-th grade level there is information available about the scores for the mathematics (MATH) and Bulgarian language and literature (BEL) exams. The histograms of these scores are presented on Figure 3 and Figure 4 respectively for the exam held for grade 4 in 2010; and on Figure 5 and Figure 6 respectively for the exam in 2011. Obviously the distribution of these parameters does not follow the normal distribution curve. The mean and the variance of the parameters are presented in Table 1. Since these two parameters are used as predictors in the regression models, the form of their distribution is not important for the quality of the model.

![Histograms of exam scores](image)

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAT 2010</td>
<td>15.9657</td>
<td>3.7116</td>
</tr>
<tr>
<td>BEL 2010</td>
<td>12.0603</td>
<td>2.6596</td>
</tr>
<tr>
<td>MAT 2011</td>
<td>15.2885</td>
<td>3.7432</td>
</tr>
<tr>
<td>BEL 2011</td>
<td>12.1160</td>
<td>2.7018</td>
</tr>
</tbody>
</table>

Table 1. The mean and the variance, grade 4
The 7-th grade level

Similarly, the distributions of the parameters, for the 7-th grade are presented on the histograms on Figure 7 and Figure 8 respectively for the exams held in 2013 and on Figure 9 and Figure 10 respectively for the exam in 2014. The mean values and the variances are presented on Table 2.

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAT 2013</td>
<td>34.1255</td>
<td>16.9015</td>
</tr>
<tr>
<td>BEL 2013</td>
<td>37.1727</td>
<td>15.0254</td>
</tr>
<tr>
<td>MAT 2014</td>
<td>22.5720</td>
<td>9.4153</td>
</tr>
<tr>
<td>BEL 2014</td>
<td>18.2325</td>
<td>5.9984</td>
</tr>
</tbody>
</table>

Table 2. The mean and the variance, grade 7
21) The distribution of these two parameters has much larger variance than expected (this is the variance if the distribution follows the normal law of data with this spread). These two parameters are considered as response variables in the regression models in the main part of the study. The large variance contributes to the small value of the determination coefficient $R^2$ (the proportion of the data explained by the model) as there is a large unexplained variance in the outcome.

**The software used**

22) The data, which should be studied in this project, are available in files with fixed width text format. This requires a preliminary processing of the data in order to combine the information from different data sources and to rearrange it in a useful dataset. Therefore the analyses of the data can be considered in two main stages.
   - Preliminary processing of the data and arranging the dataset;
   - Statistical data analyses and model confirmation.

23) The first stage is performed using Perl 5 programming language. The main advantage of this choice is that this is a very flexible script type language, designed for manipulating large text files. It is an open source. The available data files are saved as text files with “comma separated value” structure. After that the information from these files is processed and combined using a custom developed Perl program. During this process the data were checked for any type of inconsistency, which can affect the quality of the final study. As a result a well-structured dataset is produced.

24) The second stage consists of actual data analyses and model fitting. This stage can be separated in three different steps.
   - Descriptive data statistics.
   - Regression analyses for discovering the main relationships in the dataset, according to the first of the proposed models.
   - Fitting a Multilevel model to apply the second of the proposed models.

25) Taking into account the complexity of the tasks, the MATLAB system for scientific calculation was chosen. This commercial software, produced by MathWorks, is well known with its capabilities and accuracy in computation of large, multidimensional objects. Even more, MATLAB can be considered as a powerful and flexible programming language, which is important in this study, as it gives the possibility for additional data manipulations if needed. The MATLAB framework provides a very powerful “Statistical toolbox” which incorporates a large number of statistical tools and models. Using this toolbox, in order to apply the proposed theoretical models to the available data set, a custom designed program will be developed.
Description of the models

Linear Regression Model

26) The method below is based on the linear regression model, where the individual student test scores after grade 7 are regressed on the scores after grade 4, on the contextual characteristics and on the school level characteristics:

\[ G_{ij}^7 = \mu + \alpha G_{ij}^4 + \beta X_{ij} + \epsilon_{ij}, \]  

(1)

where \( G_{ij}^7 \) stands for the grade 7 results in Bulgarian language and literature, or for mathematics or total (and correspondingly \( G_{ij}^4 \) stands for the same subject grade 4 results) of the student \( i \) from school \( j \), \( X_{ij} \) is the student characteristic, included in the model (gender). Here \( \mu, \alpha, \) and \( \beta \) and the parameter to be estimated, \( \epsilon_{ij} \) are the random errors of the model, which have zero mean, are uncorrelated and normally \( N(0, \sigma^2_\epsilon) \) distributed with variance \( \text{Var} \epsilon_{ij} = \sigma^2_\epsilon \).

27) The residuals \( \epsilon_{ij} \) (the estimates of the error terms \( \epsilon_{ij} \)) from the estimated regression are used for the calculation of the value-added of the schools. More precisely, the value-added measure of the \( j \)-th school of a given subject (BEL or MATH) is the mean residual of students from this school:

\[ S_j = \frac{1}{n_j} \sum_{i \in \text{Sc}_j} \epsilon_{ij} = \frac{1}{n_j} \sum_{i \in \text{Sc}_j} (G_{ij}^7 - \hat{G}_{ij}^7). \]  

(2)

Here \( \text{Sc}_j \) is the set containing all students of the \( j \)-th school, \( n_j \) is the number of the students in this school and \( \hat{G}_{ij}^7 \) is the estimated predictor of the grade 7 exam scores of the \( i \)-th student from the estimated school.

School value-added as a random effect

28) A more general approach to the influence of the school on the students’ results is to consider the school value-added \( S_j \) as a random effect. Such models are known as mixed effect model or multilevel models. This model differs from the model in (1) by introducing an additional term – random effect \( S_j \).

\[ G_{ij}^7 = \mu + \alpha G_{ij}^4 + \beta_k X_{ij} + \gamma_l + \epsilon_{ij}, \]  

(3)

In equation (3) the terms \( \beta_k, k = 1, 2 \) and \( \gamma_l, l = 1, 2, 3, 4 \) represent the random effect of the student’s gender and the random effect of the language spoken at home, correspondingly.

29) In fact these models represent a multilevel regression model with a random intercept. In this case an important role play the variances of the random terms – the school effect \( \sigma^2_x \), the effect of the school \( \sigma^2_\beta \), the language effect \( \sigma^2_\gamma \) and the error term \( \sigma^2_\epsilon \). In Section 3.1 two different implementations of the model in (3) are presented: one without the language term \( \gamma_l \) and one with this term included.

30) Although the multilevel modeling has a number of disadvantages related to the complexity of the stochastic form and the computationally intensive procedures for its statistical estimation, it has many desirable features. One of the most important is the fact that its value added estimates “incorporate” the so called “shrinkage”. Goldstein (2011) showed that the estimated school level residuals \( S_j \) equal the raw residuals obtained from the model’s estimated fixed effects, adjusted by a shrinkage factor \( c_j \). The
constant $c_j$ is bounded by the interval $[0,1]$. Hence the residuals $S_j$ are smaller than the raw residuals. As the size of school (number of the students participating in the exam) increases, the shrinkage factor gets closer to 1. This means that for large schools the value-added residuals obtained on the basis of the OLS method and of the multilevel modelling will be similar, while for small schools the shrinkage factor has an important impact. More precisely, $S_j = \frac{n_j \sigma_j^2}{n_j \sigma_j^2 + \sigma_c^2} e_j$. In this formulae $n_j$ is the number of the students in school $j$ and $e_j$ is the mean of the raw residuals for this school: $e_j = \frac{\sum_{i \in S_j} e_{ij}}{n_j}$, where $S_j$ is the set of all students of the school $j$.

31) The factor multiplying the mean $e_j$ of the raw residuals for the $j$-th school is called “shrinkage factor”:

$$c_j = \frac{n_j \sigma_j^2}{n_j \sigma_j^2 + \sigma_c^2}$$

32) The degree of shrinkage depends on the size of the school, the increased shrinkage for a small school can be regarded as an expression of the relative lack of information in the school, so the “best” estimate sets the predicted residual closer to the overall population value, i.e. smaller schools are “shrunk” towards the national mean. In the study the shrinkage factors of schools below a certain level are calculated.

Description of the results

Estimation of the school value-added using Linear Regression Model

33) Here we consider the results, obtained using the model (1). The estimated, according to (2), school value-added for all schools in the study is reported in file result lr.xls. The model is applied for both language and math exams separately. As an estimation of the overall school value-added the average of the two specific values is used. The procedure is applied for data from 2013 and 2014 separately and together and three estimates of the added school value are obtained.

34) Let us look at the results for the combined data from years 2013 and 2014. (If we consider years 2013 and 2014 separately, similar results are obtained). The value of the estimated parameters for the models are presented in Table 3.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>SE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEL $\mu$</td>
<td>0.534</td>
<td>0.1887</td>
<td>0.0046</td>
</tr>
<tr>
<td>BEL $\alpha$</td>
<td>2.1442</td>
<td>0.0152</td>
<td>0.0000</td>
</tr>
<tr>
<td>BEL $\beta$</td>
<td>2.557</td>
<td>0.0817</td>
<td>0.0000</td>
</tr>
<tr>
<td>MAT $\mu$</td>
<td>-3.688</td>
<td>0.1668</td>
<td>0.0000</td>
</tr>
<tr>
<td>MAT $\alpha$</td>
<td>2.017</td>
<td>0.0102</td>
<td>0.0000</td>
</tr>
<tr>
<td>MAT $\beta$</td>
<td>1.0868</td>
<td>0.0766</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Table 3. Parameter estimates for the regression model

The reported $R^2$ is as low as 0.1644 (for Bulgarian language and literature) and 0.2613 (for mathematics). Histogram of the estimated residuals $e_{ij}$ for the Bulgarian language and literature and for mathematics are presented in Figure 11 and Figure 12, respectively.
35) The distribution of the schools according to the estimated school value-added in both Bulgarian language and literature and Mathematics is presented on Figure 13. The ID of the schools near the boundary of the set are labelled. Table 4 presents the estimates of the parameters by years separately and combined.

![Figure 11. Histogram of the residuals, BEL](image1)

![Figure 12. Histogram of the residuals, MATH](image2)

![Figure 13. Distribution of the schools, according to the LR model](image3)

### Table 4. Estimates of the fixed effect parameters

<table>
<thead>
<tr>
<th>year</th>
<th>parameter</th>
<th>value</th>
<th>StdErr</th>
<th>p-value</th>
<th>value</th>
<th>StdErr</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>$\mu$</td>
<td>0.9965</td>
<td>1.4722</td>
<td>0.49848</td>
<td>-2.5569</td>
<td>0.66134</td>
<td>0.0001</td>
</tr>
<tr>
<td>2013</td>
<td>$\alpha$</td>
<td>2.8414</td>
<td>0.0188</td>
<td>0.0000</td>
<td>2.1786</td>
<td>0.0157</td>
<td>0.0000</td>
</tr>
<tr>
<td>2014</td>
<td>$\mu$</td>
<td>7.0766</td>
<td>0.41978</td>
<td>0.0000</td>
<td>5.9179</td>
<td>0.2354</td>
<td>0.0001</td>
</tr>
<tr>
<td>2014</td>
<td>$\alpha$</td>
<td>0.8717</td>
<td>0.0083</td>
<td>0.0000</td>
<td>1.0453</td>
<td>0.0092</td>
<td>0.0000</td>
</tr>
<tr>
<td>XXXX*</td>
<td>$\mu$</td>
<td>5.6156</td>
<td>0.9603</td>
<td>0.0000</td>
<td>0.0069</td>
<td>0.4204</td>
<td>0.9868</td>
</tr>
<tr>
<td>XXXX*</td>
<td>$\alpha$</td>
<td>1.7223</td>
<td>0.0165</td>
<td>0.0000</td>
<td>1.7377</td>
<td>0.01074</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

* XXXX means combined data from years 2013 and 2014.

**Estimation of the school value-added as a random effect**

36) We will now apply the model (3) without the language term. It is reduced to

$$G^*_j = \mu + \alpha G^*_j + \beta + S_j + \epsilon_{ij},$$
where only the gender of the student and the school effect are treated as random effects without their interaction term. The parameter $\alpha$ represents the fixed effect of the student’s 4-grade score $G_{ij}^4$. The estimated variances of the random effect parameters and the error terms are presented in Table 5.

<table>
<thead>
<tr>
<th>Year</th>
<th>Exam</th>
<th>$\sigma_\beta^2$</th>
<th>$\sigma_\epsilon^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>BEL</td>
<td>6.8180</td>
<td>9.9046</td>
</tr>
<tr>
<td>2013</td>
<td>MAT</td>
<td>7.8192</td>
<td>11.641</td>
</tr>
<tr>
<td>2014</td>
<td>BEL</td>
<td>2.8493</td>
<td>4.4907</td>
</tr>
<tr>
<td>2014</td>
<td>MAT</td>
<td>5.2257</td>
<td>6.8165</td>
</tr>
<tr>
<td>XXXX*</td>
<td>BEL</td>
<td>4.4983</td>
<td>12.879</td>
</tr>
<tr>
<td>XXXX*</td>
<td>MAT</td>
<td>5.7376</td>
<td>11.543</td>
</tr>
</tbody>
</table>

Table 5. Estimates of the variance of the random effect and the error

* XXXX means combined data from years 2013 and 2014.

37) Histogram of the estimated residuals for Bulgarian language and literature and mathematics exams are presented in Figure 14 and Figure 15, respectively.

![Figure 14. Histogram of the residuals, BEL](image)

![Figure 15. Histogram of the residuals, MATH](image)

38) These models allow a “shrinkage” procedure, defined by equation (4), to be applied to the estimated random effect of the school. As a result the influence of the size of the school is reduced. The shrinkage factor was applied separately for two school size: 20 students and 30 students per year. The result shows that there is no meaningful difference between the results for the two values. The value 20 seems to be more appropriate, as it limits the number of classes in the school to one.

39) The distribution of the schools, according the estimated school added values (shrunken) in both Bulgarian language and literature and mathematics is presented on Figure 16. The ID of the schools near the boundary of the set is labelled. The detailed score of the school value-added for this model is presented in file result.xls.
Differences of the value-added of the schools according to the different school characteristics

40) Here we consider how the different school characteristics affect the value-added of the school. We will consider one characteristic as a particular example, which is the region in Bulgaria (defined as Region) at which the school is located. All other characteristics can be considered in the same way. Their results are presented in sections 1 to 144 in the Appendix.

41) As a measure of school value added both Normed and Scaled values are used. These two estimates of a school value-added are based on the estimation of the influence $S_i$ of the school $i$ on the students achievements, treated as a random effect in a regression model. They are computed the following way

\[
SCALED = \left[ S_i \left( \frac{100}{\max_j S_j - \min_j S_j} \right) + \left( 100 - \max_j S_j \left( \frac{100}{\max_j S_j - \min_j S_j} \right) \right) \right].
\]

\[
NORMED = \left[ \frac{\sqrt{30}}{\sigma} S_i + 100 \right].
\]

Here \([\cdot]\) means the nearest integer number and $\sigma^2$ is the estimated value of the random effect variance. The values $S_j$, $j=1,2,...,n$ comes as a result of a “shrinkage” procedure, reducing the influence of the school size on the estimated random effect.

42) The SCALED school value-added represents the position of the school in the range between 0 and 100. The value 0 means that a school has the lowest (compared to the other schools) effect of the students’ achievements and the value 100 represents the highest effect. The average SCALED effect is near 50 (approximately). The NORMED school value-added has a mean value of 100 and variance 30. The schools with higher value have a greater effect.

43) The comparison of school value-added for different school characteristics is performed for both Bulgarian language and literature (bel) and mathematics (mat) as well as for their combination, noted as XXX. Additionally a separated study is performed for different years of the exams: 2013 (X13) and 2014 (X14), as well as taking both years together (XXX). Thus for example the parameters, noted as XXX.bel.Scaled represents the school value-added, calculated using (5), based on the results of Bulgarian language and literature exam in both years together.

44) As an example, we will study the dependence of the NORMED school value-added and the region (Region) parameter, calculated for the years 2013 and 2014, and based on both mathematics and language
exams (XXX.xxx.Normed). In Table 1 of the Appendix the means of the XXX.xxx.Normed for the different values of the Region are presented.

45) In order to check the statistical significance of the observed differences in the mean values for different regions we perform one-way ANOVA with the null hypothesis stating that “There is no difference between the mean values”. Obtaining the p-value $p < 0.05$ one can reject the null hypothesis and conclude that the observed differences in the mean value of the XXX.xxx.Normed parameter are statistically significant and the region affects the school value-added.

46) Nevertheless Region is important for the mean value of the observed school value-added score, it is possible that there is no difference between two (or more) particular groups (regions). This study is presented in Table 2 of the Appendix, where the first and the second columns present the compared groups, the forth column presents the estimated difference and it is 95% confidence intervals (third and fifth columns) and the p-value for the hypothesis “There is no difference between the groups”. According to the data in this Table there is not difference between group 1 and group 5 (for example) as the p-value $p = 0.862$ is greater than the significance level 0.05. Looking at the row comparing groups 2 and 25, a p-value $p = 0.006$ can be found. This means that the observed difference -8.819 is statistically significant and schools in region 25 have a higher school value-added than the schools in region 2. This difference is between -15.534 and -1.104 with probability of 95%. This conclusion is confirmed by the figure in Section 1 in the Appendix (on page 11) where the different regions are presented as lines. The length of the line represents the variance of the score for the group. The dot represents the mean value. Two group means are significantly different if their intervals are disjoint; they are not significantly different if their intervals overlap. The same concept of study can be applied for the comparisons, presented in Section 4 of the Appendix.

**Effect of the number of teachers on the value-added of the schools**

47) This paragraph considers the effect of the number of teachers in the school to the estimated value-added. In Section 145 of the Appendix, the results of a linear regression model with predictor Number of teachers and response the corresponding school value-added are presented. The results are presented for both SCALED and NORMED scales for different year of the exams 2013 (X13) and 2014 (X14) as well as taking both years together (XXX ).

48) The conclusion is that the school value-added does not depend on the number of teachers, as the dispersion of the observations, presented in the scatter-plot is away from the confidence interval of the regression line and the distribution of the residuals is far from the Normal distribution as it can be seen on the normal probability plot. Even more the $R^2$ statistics is too low, despite that the estimation of the regression coefficients in some cases are statistically significant.

**Effect of the language spoken at home**

49) Now we will study the effect of the home language of the students. The model to be applied is (3), with the language term. The model was performed for different types of exams (Bulgarian language and literature and Mathematics) for the year 2013, as the data about the language spoken by students at home is available only for this year. The estimated values of the parameters, in this case are presented in Table 6. The estimated variances of the random effect parameters and the error terms are presented in Table 7.

<table>
<thead>
<tr>
<th>parameter</th>
<th>BEL</th>
<th>MAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>StdErr</td>
<td>p-value</td>
</tr>
<tr>
<td>$\mu$</td>
<td>-0.38387</td>
<td>2.3326</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>2.7086</td>
<td>0.018901</td>
</tr>
</tbody>
</table>

Table 6. Estimates of the fixed effect parameters

<table>
<thead>
<tr>
<th>Year</th>
<th>Exam</th>
<th>$\sigma^2_{r,\text{BEL}}$</th>
<th>$\sigma^2_{r,\text{MAT}}$</th>
<th>$\sigma^2_{\varepsilon}$</th>
<th>$\sigma^2_{\varepsilon}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>BEL</td>
<td>5.2882</td>
<td>3.1828</td>
<td>9.7986</td>
<td>11.610</td>
</tr>
<tr>
<td>2013</td>
<td>MAT</td>
<td>7.6256</td>
<td>1.9418</td>
<td>9.7986</td>
<td>11.610</td>
</tr>
</tbody>
</table>

Table 7. Estimates of the variances of the random effect and error

16
The values of the estimated terms for the different languages are presented in Table 8.

<table>
<thead>
<tr>
<th>Language ID</th>
<th>BEL</th>
<th>MAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Bulgarian</td>
<td>4.5739</td>
<td>3.0105</td>
</tr>
<tr>
<td>2. Roma</td>
<td>-0.7146</td>
<td>-0.3495</td>
</tr>
<tr>
<td>3. Turkish</td>
<td>-3.6203</td>
<td>-1.6827</td>
</tr>
<tr>
<td>4. Other</td>
<td>-0.2389</td>
<td>-0.9783</td>
</tr>
</tbody>
</table>

Table 8. Values of the estimated terms for the different language

The information provided in Table 8 shows that including the language term will most likely lead to increasing the value added measures for the schools with larger number of students speaking non-Bulgarian language at home.

50) Histograms of the estimated residuals $\epsilon_{ij}$ for the Bulgarian language and literature and mathematics exams are presented in Figure 17 and Figure 18, respectively.

51) This models allow a “shrinkage” procedure, defined by equation (4), to be applied to the estimated random effect of the school. As a result the influence of the size of the school is reduced. Again the shrinkage factor can be applied on the schools with less than 20 students. The distribution of the schools, according to the estimated school value-added (shrunked) in both Bulgarian language and literature and Mathematics, is presented on Figure 19. The IDs of the schools near the boundary of the set are labeled.
Key Findings and Recommendations

52) This study considers how different statistical models can be applied for the estimation of school value added in Bulgarian schools using the data available at the moment. Three different approaches are performed and the results show that all of them can be applied for estimation the school value-added. The available data consists of exams for 2013 and 2014, with different numbers of meaningful indicators. School value-added was estimated using two parallel models, one for Bulgarian language and literature and one for Mathematics, and the compound estimation of the school value-added was calculated as an average of these two measures. For the purpose of the study the two measures of school value-added are treated separately when the characteristics of the models and the relationships are considered.

53) First a model, based on the linear regression, was considered. The main disadvantage of this model is the problem how a schools with small number of students should be treated. Other problem that arises from this model is the assumption that the school affects all the students in one and the same way. This model results in a relatively small value of $R^2$ statistics and a distribution of residuals which differs from Normal. The plots show a relatively high proportion of observations with a relatively high value of the residuals. This leads to the conclusion that there exists a sub-population with a relatively low performance on the exams.

54) The second model treats the school value-added as a random effect. The positives of this approach are the assumption that the school affects the students in a different way and that on the school level the average effect is observed. In addition this approach gives the opportunity to treat the small size schools with the so called shrinkage factor, reducing the effect of the small sample sizes. This model also shows a bi-modal distribution of the residuals in the case of Bulgarian language and literature exam, but in the case of mathematics exam the residuals follow the normal distribution.

55) In the third model, an additional random effect term is introduced. This term represents the student’s language spoken at home. This data are available only for exams held in 2013, so the model is applied only for this year. The results shows an approximately normal distribution of the residuals for both exams and a positive correlation between the school value-added for BEL and MAT.

56) The results from the analysis allow the following conclusions:

57) The linear regression model assumes that every school has equal and fixed effect on the students who learn in it. The value-added of the school is the value of this effect. The value-added of the schools with small number of students in the population is strongly influenced by the results of individual students who have relatively high (or low) results, i.e. one can expect a relatively high error in the estimation of the value-added of the small schools. In this model, there is not a methodologically correct approach to this problem. Another disadvantage of the model is the impossibility to handle the non-negligible number of unexplained data with huge variation, as well as the set of observations with relatively low scores (a relatively large proportion of observations follow high positive error).

58) Some of the above mentioned disadvantages could be solved using the multilevel regression model with random effects. This model assumes that the school has different effect on different students. The value-added of the school is the mean of these effects. Although this model applied to the available data demonstrates the same problem with the existence of subpopulations with relatively low scores, it allows a methodologically proper treatment of the schools with a small number of students, which makes this method more appropriate.

59) The random effect model can be elaborated by including the effect of the language spoken at home. Even though the language data are available only for the cohort that took exams after grade 7 in 2013, the model shows a good fit, because it appropriately models the subpopulation with relatively low scores. As a result, the above-mentioned relatively high error for the students with low scores. This model allows a shrinkage treatment of the small schools and has an additional advantage, described below.
The two random effect models described above are applied using the results of the exams of two subjects: Bulgarian language and literature (BEL) and mathematics (MAT). This way two separate value-added estimates (one for BEL and one for MAT) for each school are obtained. According to the results, the random effect model without the language term does not allow a sensible interpretation of a unique combined value-added (i.e. the mean of BEL and MAT value-added). This is because there are many schools with relatively high BEL value-added and relatively low MAT value-added and vice versa. In this case a two-dimensional representation as Figures 16 and 19 seem to be more appropriate.

The random effect model with language term, however, shows a positive correlation between the BEL and MAT value added. This is why in this case a combined value-added (i.e. the mean of BEL and MAT value-added) might be considered as a composite indicator or the value-added of the school.

Based on the above discussions of the characteristics and the properties of the piloted modes on the available data the following recommendation can be made: the random effect model with language term is the most appropriate for the estimation of the school value-added, even though this model is applied only to the population that took grade 7 exam in 2013. Given the imprecision of the estimations produced by the statistical models, a viable option is to introduce “confidence” intervals or broader quality bands for presentation of the value added results, similar to the approach that has been adopted in Poland.
References


