

Variability Analysis for Effectiveness and Improvement in Classrooms and Schools in Upper Secondary Education in Slovenia: Assessment of/for Learning Analytic Tool

D. Zupanc^{1*}, M. Urank¹ & M. Bren²

¹National Examinations Centre, Slovenia

²University of Maribor, Slovenia

From 1995 data on students' achievement in schools (i.e. teacher's grade) and all data on achievement in the five-subject group certificate – *Matura* exam – have been systematically gathered for the entire yearly cohort of students in the upper secondary education in Slovenia. The presentation describes an on-line data selection system and data analysis tool designed for national subject testing committees, schools and teachers as feedback for analyses of effectiveness, for the improvement of teaching in the classroom and for better learning. The use of data to inform teacher planning of future instructional activities for new groups of student (or for adjustment of policies) in the long run can lead to systematic improvement. That might be considered as a secondary level of formative assessment - a key step in a continuous PDCA (Plan, Do, Check, Act) process as well as in establishing and assuring the quality of education.

Keywords: *Gathering data; Data Analysis Tool; Connecting Formative and Summative Assessment; Secondary Level of Formative Assessment; Classroom Effectiveness;*

Introduction

Countries in Central and Eastern Europe had uniform national systems with controlled input. Curricula, timetables, learning content and methods were prescribed. A single textbook was approved for any individual subject in each grade. Outcomes were rarely externally assessed. Previously, in most countries in this part of Europe no data on students' achievement were gathered in school systems. The majority of countries had uniform state or regional school systems where all the inputs for all subjects in all schools were centrally planned and carried out in all classrooms in a uniform manner; the outputs, however, were seldom systematically evaluated. It was expected that once the necessary inputs have been provided the desired outputs would flow naturally (Fertig, 2000). Slovenia was one of those countries. As a developing country Slovenia has endeavoured to catch up with and join the developed countries. The source of change in the education system in Slovenia can be traced back to the development of examinations at the conclusion of secondary education. In accordance with OECD directives in 1995 Slovenia launched examination reforms and introduced the so-called external *Matura* examination: a centrally administered school-leaving examination at the end of Upper Secondary Education (USE) consisting of five units (Gabršček & Bethell, 1996). This was a new approach, the first in any of the central or eastern European countries (Gabršček, 2001).

What happens in many public examination systems is that the criterion information is ignored and the results are used only normatively to rank students for the award of grades and for

*Corresponding author. National Examinations Centre, Ob železnici 16, 1000 Ljubljana, Slovenia. Tel +386 1 5484616. Fax: +386 1 5484601. E-mail:darko.zupanc@guest.arnes.si

selection into post-examination study options (McGaw, 2006). The potential of data gathered to be used as feedback for improvement, more effective teaching and teachers' work, the community of teachers (learning community), schools and on the national level, however, has remained practically unused. Interest in the use of examination results to monitor the effectiveness of schools and to hold them responsible for their performance is increasing (Crighton, 2003).

Some countries that do not now have external examinations and/or inspection systems are considering adopting such approaches to ensure greater school accountability. By contrast, a few countries that have promoted external examinations are paying greater attention to the potential for school-based evaluation to shape school improvement. Policy officials can learn much from the experiences of their counterparts. No matter which approach is chosen, assessment and evaluation are only really effective if the data gathered at different levels are taken into account throughout the systems (Formative assessment, 2005).

Like Fullan (2000: 470) recommends, the goal is to establish a data-gathering system aimed at developing "assessment literacy" on the part of local and regional groups. Ideally, information gathered in assessment and evaluation is used to shape strategies for improvement at each level of the education system. At the classroom level, teachers gather information on student understanding, and adjust teaching to meet identified learning needs. At the school level, school leaders use information to identify areas of strengths and weaknesses across the school, and to develop strategies for improvement. At the policy level, officials use information gathered through national or regional tests, or through monitoring of school performance, to guide investments in training and support for schools and teachers, or to set broad priorities for education. In this way, summative information is used formatively at each level of the system. Teachers, school leaders and policy officials are more likely to use assessment information when assessment is well coordinated, and it is clear why and how the information is relevant to their work (*Formative assessment*, 2005).

External assessment is unlikely to tell teachers much that is new about differences in performance levels among students in their own classes. What they can add is information on how that class is performing in relation to others in similar schools elsewhere or in dissimilar schools and assessment should become a mechanism for improvement (McGaw, 2006: 4).

The major barriers to wider practice of formative assessment that emerged from the OECD studies include (McGaw, 2006; *Formative assessment*, 2005):

- the tension between classroom-based formative assessment of student learning, and high visibility summative tests and
- a lack of connection between systemic, school and classroom approaches to assessment and evaluation.

Too often, the information gathered through national or regional monitoring systems, or even in school-based evaluations, is seen as irrelevant or unhelpful for teaching. Too often, the information gathered in classrooms is seen as irrelevant for policy making.

Most of the studies tended to be focused more upon gathering "process" data at the level of the school rather than at the level of the classroom, where instruction is centred (Sheerens, & Bosker, 1997). Policy often ignores classroom level variables, to its detriment. As Reynolds (1998) points out, the greatest variations in student learning occur not among schools, but within schools, among subject departments and individual teachers. Research results highlighted the influence of teachers and classrooms on students learning, attainments and progress (Burstein, 1980; Cronbach, Linn, Brennan, & Haertel, 1997; Guion, 1995; Kyriakides, Campbell, & Gagatsis, 2000; Lee, Bryk, & Smith, 1993; Muijs, & Reynolds, 2000; Webster, & Fisher, 2000). This implies that it is more important to focus on classroom variables than on school variables. Creemers (1994) argues that the classroom is the most important place for achieving educational effectiveness.

The fact that within-school variation by classroom in learning gains varies widely, as the study by Kyriakides, Campbell and Gagatsis study suggested (2000), is an issue which needs further exploration. Only few of the studies have looked at both school and classroom effectiveness simultaneously (Mortimore, Sammons, Stoll, Lewis, & Ecob, 1988).

In the USA the Tennessee Value Added Assessment System (TVAASS) was developed to provide unbiased estimates of the influences that school systems, schools, and teachers have on academic gains of students in a number of subjects. To support this process, the largest longitudinally merged database of student achievement data in the U.S. has been constructed. Applying the TVAASS statistical methodology to this database via software developed to handle the size and scope of the process, indicators of system, school and teachers effectiveness have been provided to the educational community in Tennessee for the past several years. (Topping, & Sanders, 2000)

The Recommendation of the European Parliament and of the Council on European cooperation with regard to the qualitative assessment of school education (*Qualitative assessment of school education*, 2001) highlights the fact that exchange of information constitutes the bulk of the Community added value. Information exchange should be focused in particular on methodological developments and examples of good practice, especially through the use of new technologies, data collection, and by developing tools. The aims are to:

- create an Internet database for the dissemination of tools, examples of good practice, and effective tools for assessing the quality educational establishments of as feedback for improvement
- make an inventory of the instruments and strategies for quality assessment in primary and secondary education already in use in the various Member States;
- incorporate the results obtained into existing programmes and develop the existing networks;

Outcome-based indicators are used for international comparisons of efficiency of school systems between countries where there is important variability in achievement within a country. In a similar manner, students' achievement in schools and variability in achievement can be used for analyses of efficiency of schools. Basic units of teaching and learning within schools are classrooms, and analyses of achievement and variability at the classroom level are a rich source of information for establishing efficiency as well a sound basis for improvement in education.

In Slovenia, a tool has been developed and the basic data in the database are derived from school (i.e. teachers') assessment and from external examinations which are held for all students who completed USE. The database comprises the existing data already gathered in the school system for all subjects where knowledge is externally assessed on completion of USE and overall achievement is calculated. We investigated students' achievement in the wider range of the national curriculum (overall achievement, compulsory subjects and also optional subjects), not exclusively in English and mathematics (Kyriakides, Campbell, & Gagatsis, 2000). Analyses of students' achievement and its presentation are comparable at the national, school as well as classroom levels for several consecutive years, from 2002 until 2006. The aim is to balance summative and formative assessments and to better link assessment and evaluation at each stage of the system. Thus, data should not be used solely for summative evaluation with normative-referenced interpretation (*assessment of learning*), but as a tool for formative evaluation with criterion-referenced interpretation (*assessment for learning*) where results would contribute to more effective teaching and learning. The tool makes it possible for subject experts, head teachers and also teachers in classrooms to compare 'their' achievement and also allows for different variability analyses which form the

basis for the interpretation of achievement, discussions on efficiency as well as potential improvement.

Data, Instruments and Methods

Nowadays, two major types of USE in Slovenia include 80% of each generation: 40% in General Secondary Education (GSE) and 41% in Secondary Technical Education (STE). In addition, in the 2006/07 school year 17% of all students are enrolled in Secondary Vocational and 2% in Short-term Vocational programmes. At the end of four years of GSE, candidates complete secondary studies with the *Matura* certificate. At the end of four or five years of STE, candidates complete secondary studies with the Vocational *Matura* certificate.

The paper describes an on-line data selection system and data analysis tool designed for national subject testing committees, schools and classroom teachers as feedback for analyses of effectiveness, for the improvement of teaching in the classroom and for better learning. Data on students' achievement (i.e. teacher's grade) and all data on achievement in the five-subject group certificate *Matura* (and all data on achievement in the four-subject group certificate Vocational *Matura*) have been systematically gathered for the entire yearly cohort of students. Data have been gathered for certificates at the end of USE as well as for selection procedures for enrolment in tertiary education.

Data on students' achievement at the end of schooling and for *Matura* have been gathered for the last five years: $5 \times 2 = 10$ exam sessions = 55,000 candidates, or 45,000 of those who sat for all five subjects of *Matura* for the first time. Data are gathered for $45,000 \times 5 = 225,000$ *Matura* exams. Data on students' achievement at the end of schooling and for Vocational *Matura* are gathered too; in the last five years: $5 \times 3 = 15$ exam sessions = 85,000 candidates, or 65,000 of those who sat for all four subjects of Vocational *Matura* for the first time. Data are gathered for $65,000 \times 4 = 260,000$ Vocational *Matura* exams. Data include the achievement of the entire yearly cohort, i.e. 110,000 secondary school students in the last five years, from 2002 until 2006. This represents more than 110,000 of young citizens of Slovenia who sat for both types of *Matura* exam for the first time, which represents more than 5.5% of the entire Slovene population: $225,000 + 260,000 = 485,000$ exams, i.e. almost 0.5 million exams!

Assessment of/for Learning Analytic Tool as Instrument

The Qualifications and Curriculum Authority (QCA) in the UK defines Assessment for Learning, or formative assessment, as (Glossary, 2006)

“the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there.”

The National Examinations Centre with the support of the Slovene Ministry of Education and Sport and the European Social Fund has developed the Assessment of/for Learning Analytic Tool (ALAT Tool). Teachers and head teachers in Slovenia and other professionals will be able to access examination database, analyse assessment results, interpret achievement of their students, and analyse efficiency of teaching and learning in secondary schools. Fullan (2000) said that one of the strategies in redefining accountability is that teachers and head teachers become more "assessment literate" and assessment becomes more transparent.

The main feature of the ALAT Tool is a searchable database that makes it possible to:

- compare students' achievement to peer groups outside school;
- compare students' achievement to classes within a school;
- analyse possible drop-out rates in different programmes, schools and classes;
- analyse teacher grading and external marking;

- analyse students' selection of different optional subjects in classrooms and schools and selection of different tiers of compulsory subject exams;
- compare outcomes in one subject with another and with overall achievement (grade);
- analyse students' outcomes in different kinds of assessment: written, oral, coursework, etc;
- recognise trends;
- identify (diagnose) strengths and weaknesses in learning concepts to shape teaching for improvement;
- analyse achieved objectives and performance standards with items analysis.

The ALAT Tool allows access to data for seven types of analyses and for different combinations between them:

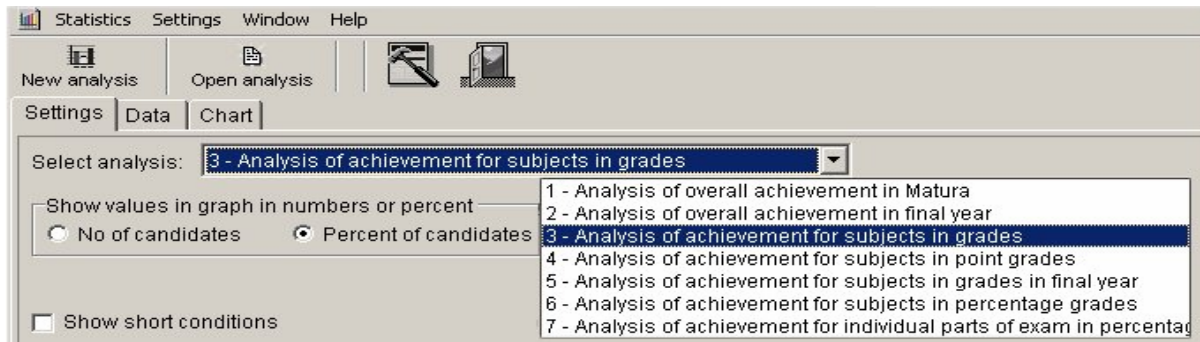


Figure 1. Types of analyses and their combinations in the ALAT Tool

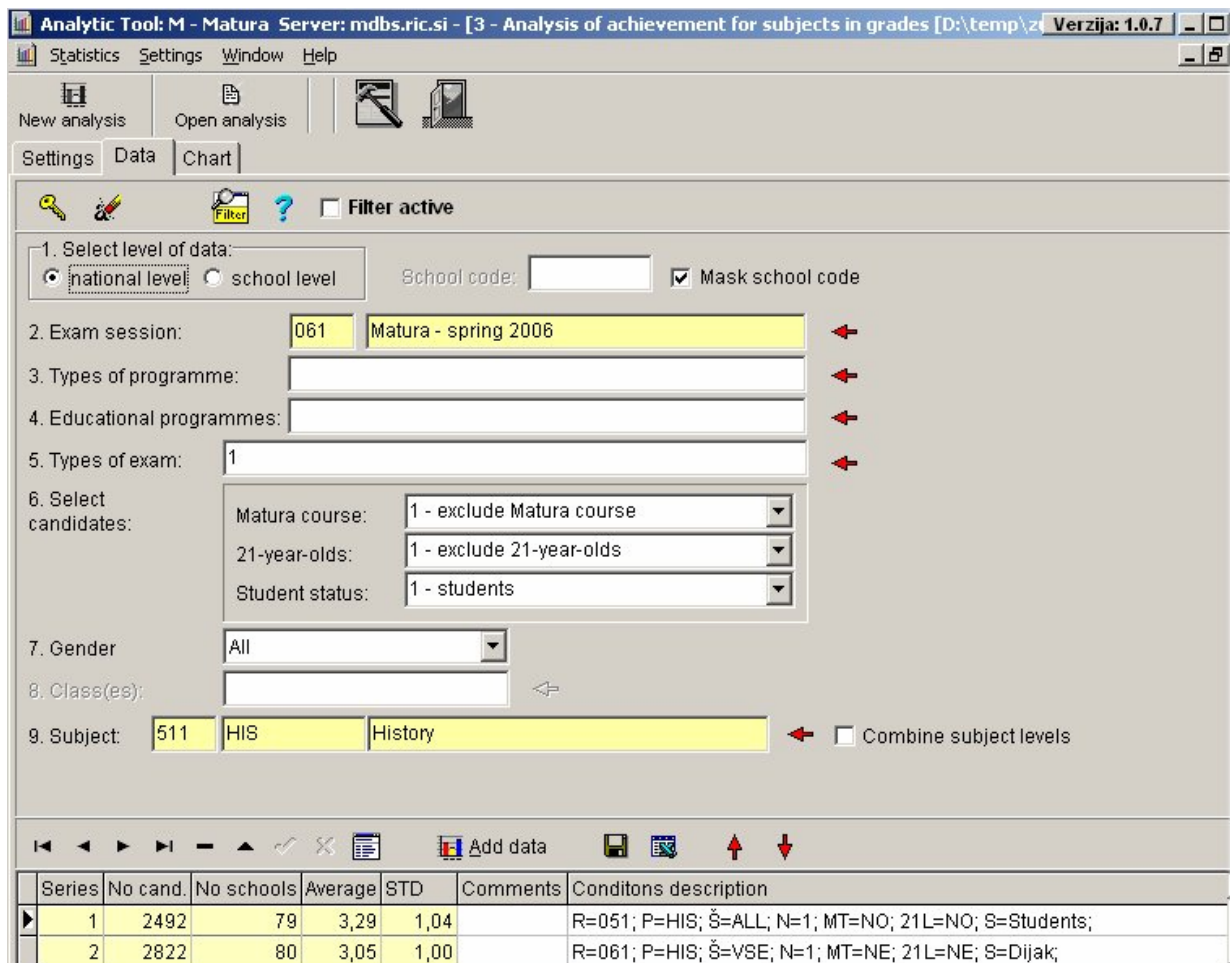
- 1) Analysis of overall achievement in *Matura*. In *Matura*, each candidate gets a grade for each of the five subjects, from 1 (Insufficient, i.e. Fail) to 5 (Excellent) at Foundation tier, and from 1 to 8 at Higher tier. Candidates pass *Matura*, if they get at least grade 2 (Sufficient) in all subjects (there is an exception, i.e. a Near Pass grade). Successful candidates with the lowest overall achievement get 10 points, while the overall achievement of the best candidates, i.e. those taking three subjects at Higher tier ($3 \times 8 = 24$) and two at Foundation tier ($2 \times 5 = 10$), is 34 points.
- 2) Analysis of overall achievement in final year. Prior to taking *Matura*, candidates must successfully complete secondary education, i.e. they have to get at least grade 2 (or higher) in school subjects. On completion of secondary education (i.e. prior to taking *Matura*), each candidate's overall achievement is determined on the school level. The scale extends from 1 (Insufficient), which means that a candidate cannot take *Matura*, to 5 (Excellent), the highest overall achievement (the passing grade being 2, Sufficient).
- 3) Analysis of achievement for subjects in grades. The traditional marking scale for subjects at Foundation tier in *Matura* has five classes, from 1 (Insufficient) to 5 (Excellent). In subjects at Higher tier, with an eight-class marking scale, each grade is converted into the traditional five-class scale so that the same subject at Foundation and Higher tiers can be compared against the same marking scale.
- 4) Analysis of achievement for subjects in point grades. Knowledge in subjects at Higher tier is assessed on the scale from 1 to 8, and in subjects at Foundation tier from 1 to 5, i.e. at Foundation tier point grades not differ from traditional grades.
- 5) Analysis of achievement for subjects in grades in final year. Prior to taking *Matura*, each candidate gets a teacher's grade for compulsory as well as optional *Matura* subjects in the school in the final year of secondary education. School grades range from 1 to 5; only candidates with a 'positive' grade in school, i.e. from 2 to 5, are allowed to take *Matura*.
- 6) Analysis of achievement for subjects in per cent. Achievement in all parts of *Matura* exam is assessed in raw scores, while syllabuses determine the proportions of written, i.e. external, as well as school-based parts of the exam (i.e. oral exams and different types of

course work, e.g. written assignments, laboratory work, etc) so that together they make 100%. Before a grade is determined on the 5- or 8-class scale, a candidate's achievement in each of the exams is expressed in percent. Each year benchmarks (for the conversion of percent into numerical grades) are set according to a special procedure.

- 7) Analysis of achievement for individual parts of exam in per cent. In all *Matura* subjects, the proportions of the written, i.e. external, and the school-based parts are usually 80% (or 75%) and 20% (or 25%) respectively. They are determined in syllabuses. A school-based part of an exam can be an oral exam, a written assignment, laboratory work, etc. Databases contain information on each candidate's achievement for each subject, separately for the written, i.e. external, part of *Matura* and the school-based part, the assessment of which is carried out in schools (by the candidate's teacher of the subject in question or by the examination board, where the candidate's teacher is the chief examiner).

The power of the instrument lies also in the fact that analyses can be freely combined, where it makes sense of course. Sensible combinations of analyses are for example: 1 – Analysis of overall achievement in *Matura* and 2 – Analysis of overall achievement in final year; or 3 – Analysis of achievement for subjects in grades and 5 – Analysis of achievement for subjects in grades in final year or 7 – Analysis of achievement for individual parts of exam in per cent, etc.

It is necessary to narrow down the multitude of data gathered in the system, i.e. to determine the (sub)group of candidates to be dealt with. Selections can be made in nine directions, or steps.



The screenshot shows the 'Analytic Tool: M - Matura' software interface. The title bar indicates the server is 'mdb.s.ric.si' and the analysis is for 'grades [D:\temp\z]'. The version is 'Verzija: 1.0.7'. The interface includes a menu bar (Statistics, Settings, Window, Help) and a toolbar with icons for 'New analysis' and 'Open analysis'. Below the toolbar are tabs for 'Settings', 'Data', and 'Chart'. A 'Filter active' checkbox is present. The main area contains nine numbered selection steps:

- Select level of data: national level, school level. School code: [text box]. Mask school code.
- Exam session: 061, Matura - spring 2006.
- Types of programme: [text box].
- Educational programmes: [text box].
- Types of exam: 1.
- Select candidates:
 - Matura course: 1 - exclude Matura course
 - 21-year-olds: 1 - exclude 21-year-olds
 - Student status: 1 - students
- Gender: All.
- Class(es): [text box].
- Subject: 511, HIS, History. Combine subject levels.

At the bottom, there is a data table with columns: Series, No cand., No schools, Average, STD, Comments, and Conditions description.

Series	No cand.	No schools	Average	STD	Comments	Conditions description
1	2492	79	3,29	1,04		R=051; P=HIS; Š=ALL; N=1; MT=NO; 21L=NO; S=Students;
2	2822	80	3,05	1,00		R=061; P=HIS; Š=VSE; N=1; MT=NE; 21L=NE; S=Dijak;

Figure 2. Selections in narrowing down the multitude of data

Step 1: The level is selected: either national, where analyses will be performed on data for the entire yearly cohort of students, or school, where analyses will be performed on data for students of an individual school or a group of students within a school. Each school has its code. Users in schools can access all data at the national level and only data for their own school at the school and classroom levels. Asterisks are used to mask the school's code.

Step 2: Exam session is selected. There are two exam sessions of *Matura* each year, i.e. $2 \times 5 = 10$ exam sessions in the last five years, from 2002 until 2006. There are three exam sessions of Vocational *Matura* each year, i.e. $3 \times 5 = 15$ exam sessions.

Step 3: Type of programme is selected so that schools can compare achievement of their students with their peers. GSE in Slovenia has different streams: General, Technical, Classical, Arts, as well as the specific *Matura* Course. Vocational *Matura*, however, is taken by students from STE, Vocational Technical Education (VTE), the Vocational Course, etc.

Step 4: Educational programme is selected. All secondary educational programmes leading to *Matura* are listed systematically. Especially varied is STE leading to the Vocational *Matura* certificate. These are different programmes from the area of electrical engineering, electronics, mechanical engineering, economy, tourism, etc. The purpose of this selection is to allow for a comparison of achievement in *Matura* in an individual school or in different classes of a school only with achievement of their peers in the same type of educational programmes in Slovenia.

Step 5: Type of exam is selected. Candidates take *Matura* in different ways. The majority sit the entire (i.e. all five subjects) *Matura* for the first time; some take *Matura* for the second/third time because they failed the first time (have to re-sit one or several of *Matura* exams) or because they want to improve a grade. The purpose of this selection is similar, i.e. to allow for a comparison of only comparable groups.

Step 6: Candidates are selected. Some subgroups of candidates in *Matura* attended part-time programmes prior to taking *Matura* and their achievement significantly differs from the most common group of full-time students who take *Matura* after completing USE. When analysing the data, these special groups can be either included in or excluded from the analysis, or else only such a group can be analysed. These are candidates taking *Matura* after completing the one-year *Matura* Course or the Vocational Course, and candidates who are allowed to take *Matura* at the age of 21 or more without having successfully completed secondary education. This step also allows combinations of subgroups of full-time students and/or adult candidates.

Step 7: Gender is selected. Candidates' gender can be selected: both genders together or boys and girls separately.

Step 8: Class is selected. During the educational process, candidates of a school (prior to *Matura*) attend one of the several Year 4 classes, e.g. 4A, 4B, /.../ 4H. For the analysis of the data, a single class or several classes of a school can be selected – classes taught by the same teacher can be grouped together.

Step 9: Subject is selected. When analysing achievement in subjects, an individual subject is selected. There are 33 *Matura* subjects: some exams are prepared and administered at Foundation and Higher tiers. The ALAT Tool will allow for the selection of a specific group of students, which at the national level is not necessarily linked to one school or one class within a school; e.g. it will be possible to analyse achievement of a group of students taking *Matura* in one optional subject irrespective of the class of their school they attended. It will be possible to perform other analyses for such a group of students: for exams in other subjects taken either in *Matura* or in school.

When users of the ALAT Tool select a (sub)group, they make an on-line enquiry on the central server of the National Examinations Centre, where the number of students of the (sub)group is calculated on the achievement of the selected group of candidates (i.e. the

number, point average, standard deviation and other information necessary for the representation of the distribution of grades, points, etc).

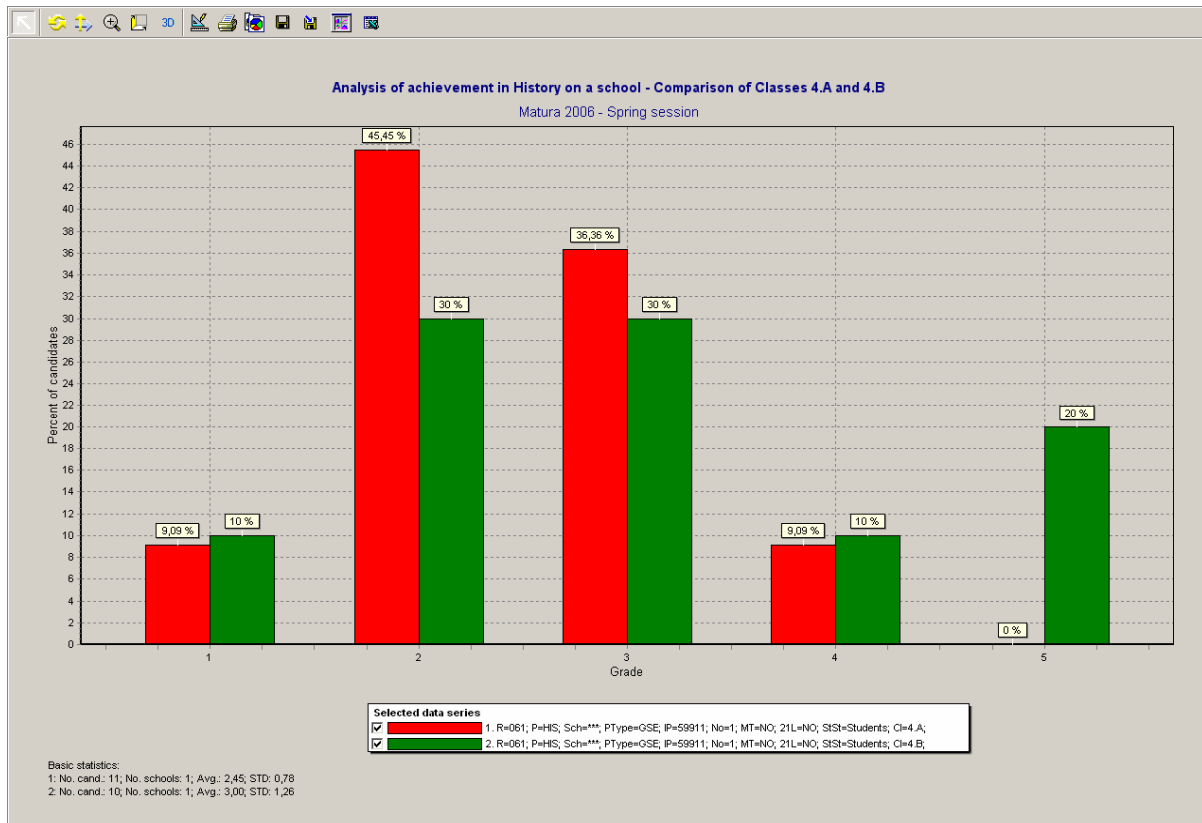


Figure 3. Analysis of achievement in History in a Slovene school: Comparison of Classes 4A and 4B (*Matura* 2006, Spring session)

The ALAT Tool enables the representation of the distribution and allows for a comparison of several distributions with different types of graph: column or line graphs, pie charts, etc. Data from each of the seven types of analysis and their combinations can be exported to Microsoft Excel program for further analysis.

Figure 4 shows the distribution of grades in History in *Matura* (on the scale from 1 to 5) for two classes of the same school for the Spring session of *Matura* 2006.

Variability in achievement between schools, different classes of the same school, an individual class and the school, and different subjects and years can be analysed. Analyses show variability and deviations in achievement at different levels of the school system, national, school and classroom.

We look at the classroom, school and national levels simultaneously: the gathered data are the same and analyses are comparable. The information gathered at each level of the system can be used to identify strengths and weaknesses and to develop strategies for improvement.

Results of variability analyses

Some analyses at the national, school and classroom levels will be reported. The paper covers the analysis of variability of achievement in one educational programme, i.e. the one leading all candidates to the same certificate and the possibility to enrol in university education; thus being the most comprehensive (ca 6,000 candidates a year, or 30,000 in five years).

Analysis: Proportion of successful students in upper secondary schools

At the national level, experts as well as parents and the wider public often ask themselves how successful schools really are. Like it or not, this question is often simplified to mean *What proportion of students pass Matura after having first completed USE?* Figure 4 shows the analysis of the proportion of successful students from 53 upper secondary schools (USS) in one exam session, i.e. the Spring session of *Matura* 2006: 11 schools, i.e. 21%, out of 53 general USSs had 100% of successful students; on the other hand, there were 'only' 10 schools, i.e. 19%, with a proportion of successful students below 95%. The large majority of these schools, i.e. ca 60%, have between 95% and 100% of successful students. So in this respect, with the exception of the 10 mentioned schools, variability in the proportion of successful students is not great. Variability between USSs in the proportion of successful students according to the criterion Passed/Sat is not surprising.

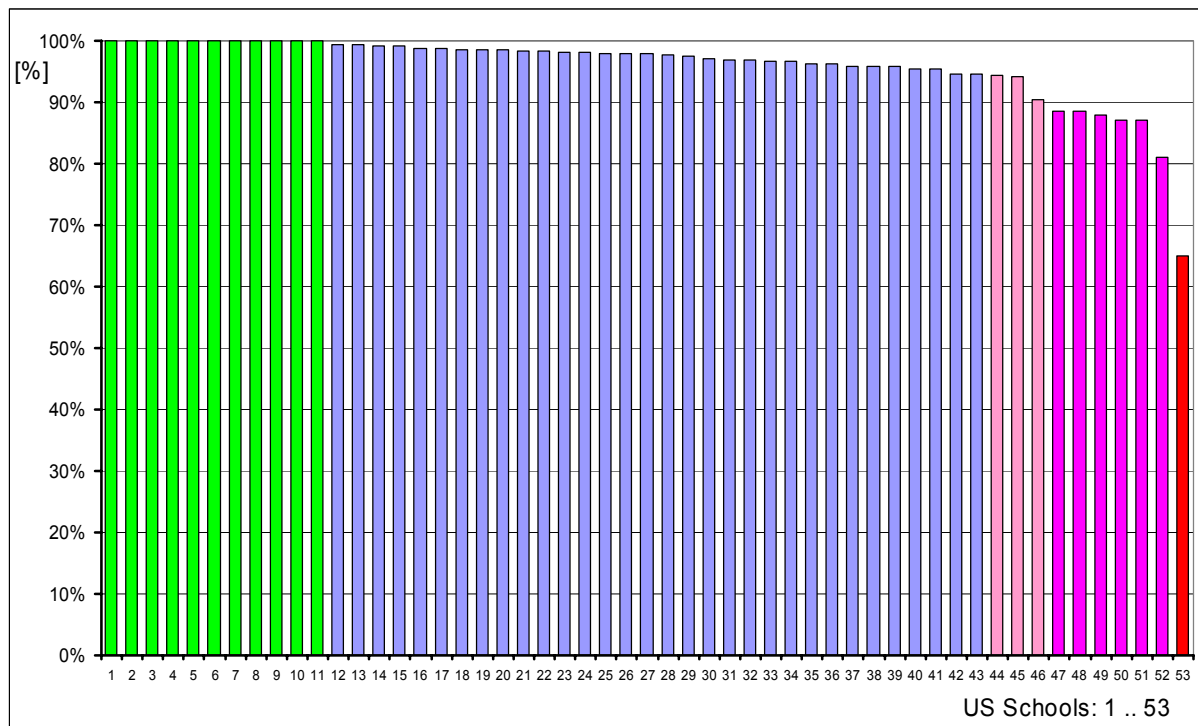


Figure 4. Proportion of successful students in USSs in *Matura* 2006 (Spring session)

What are the differences between schools if proportions are shown according to the criterion? If one, however, is more familiar with the Slovene system of education, the above question can be formulated differently. Data on all candidates entering for *Matura* are gathered in the system before the school year is over. Candidates can take *Matura* only if they have successfully completed the final year of USE. The decision, however, depends on everyone in the school: apart from students, teachers and the policy of school. The proportion of successful students, i.e. those who pass *Matura* as compared to those who take *Matura*, can thus significantly change if achievement is considered with regard to all students who were preparing for *Matura* in the final year of USE.

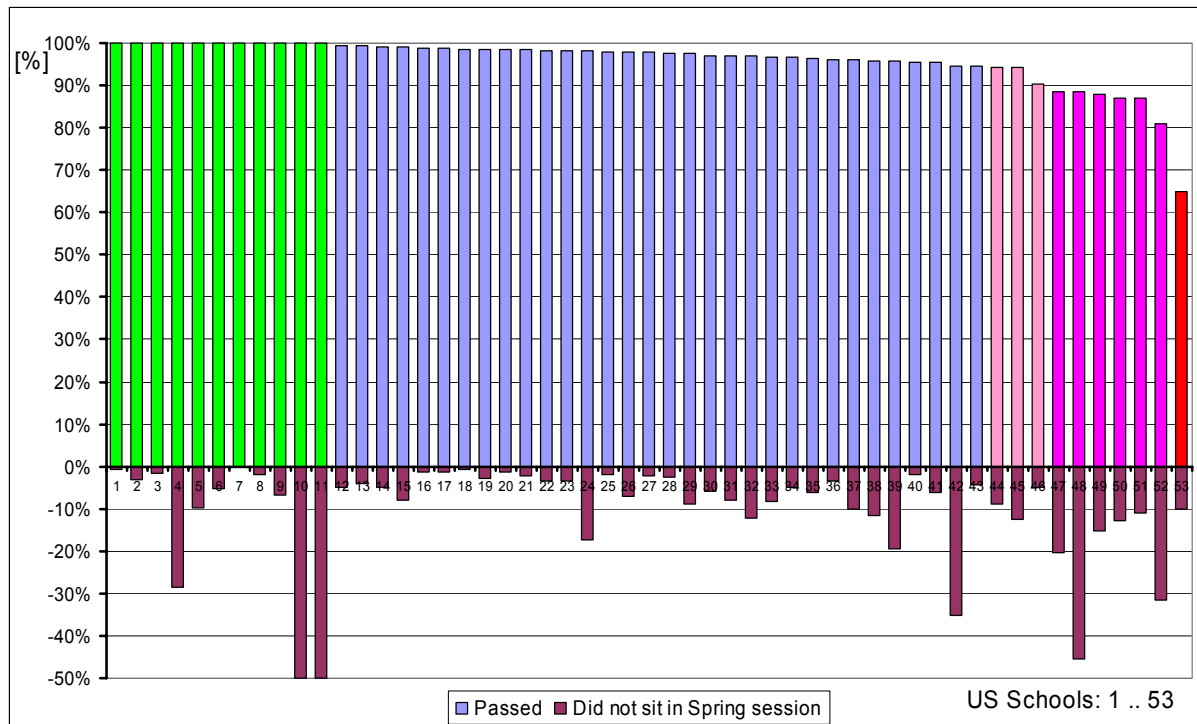


Figure 5. Proportions of successful students and those who did not sit the exam in USSs (*Matura* 2006, Spring session)

Proportions of successful students (of those that took *Matura*) in schools are shown in the upper part of the histogram in Figure 5. The upper part is the same as in Figure 4. Proportions of the number of candidates, who took preparatory lessons for the Spring session and consequently applied for the exam (2006) but did not take it, are shown in the lower part. The lower part in fact shows the proportion of students in the final year in individual USSs, who were unsuccessful in school, i.e. failed in one or several subjects. The general picture can be, to a lesser extent though, deformed by candidates who did not take *Matura* for other reasons, i.e. they changed schools a month or two before the exam, they took the exam in two sessions, etc.

In this case, variability between schools is much greater. There are cases where school grades do not prevent any of the students to take the exam in the Spring session, cases where there are a few per cent of such students in a school, and cases where a few tens per cent of students fail to successfully complete the final year of USE; proportions sometimes comprise up to one half of all students from the final year prior to *Matura*. Such differences between schools as revealed through analyses by all means deserve consideration, looking for causes, adequate interpretations and most likely induce future changes. National analyses are available to all practitioners, head teachers and teachers in classrooms, too, to be able to compare, analyse and evaluate their efficiency (*Indicators on the quality of school education*, 2000). The point is that this is the matter for debate in schools, the basis for their (self-)evaluation so that they look for reasons and explanations and suggest future improvement.

Analysis: Differences in the selection of subjects

Slovene *Matura* is an exam consisting of five subjects, three are compulsory, i.e. the mother tongue (for the majority of candidates this is Slovene), Maths and a foreign language, and two are optional, i.e. a candidate can select two from a given list of subjects. The legitimate question is whether there is a significant difference between groups of students who select certain optional subjects. This is important for subject experts as well as for selection

procedures for enrolment in tertiary education. In a system with a five-subject group certificate, such as *Matura*, the common ground for a comparison of groups of students who selected a certain optional subject can be candidates' achievement in the three compulsory subjects. In Slovene, the highest grade that can be achieved in *Matura* is 8; this is also the case in Maths and the foreign language if the exam is taken at Higher tier. Consequently, a candidate can get up to $3 \times 8 = 24$ points in the three compulsory subjects.

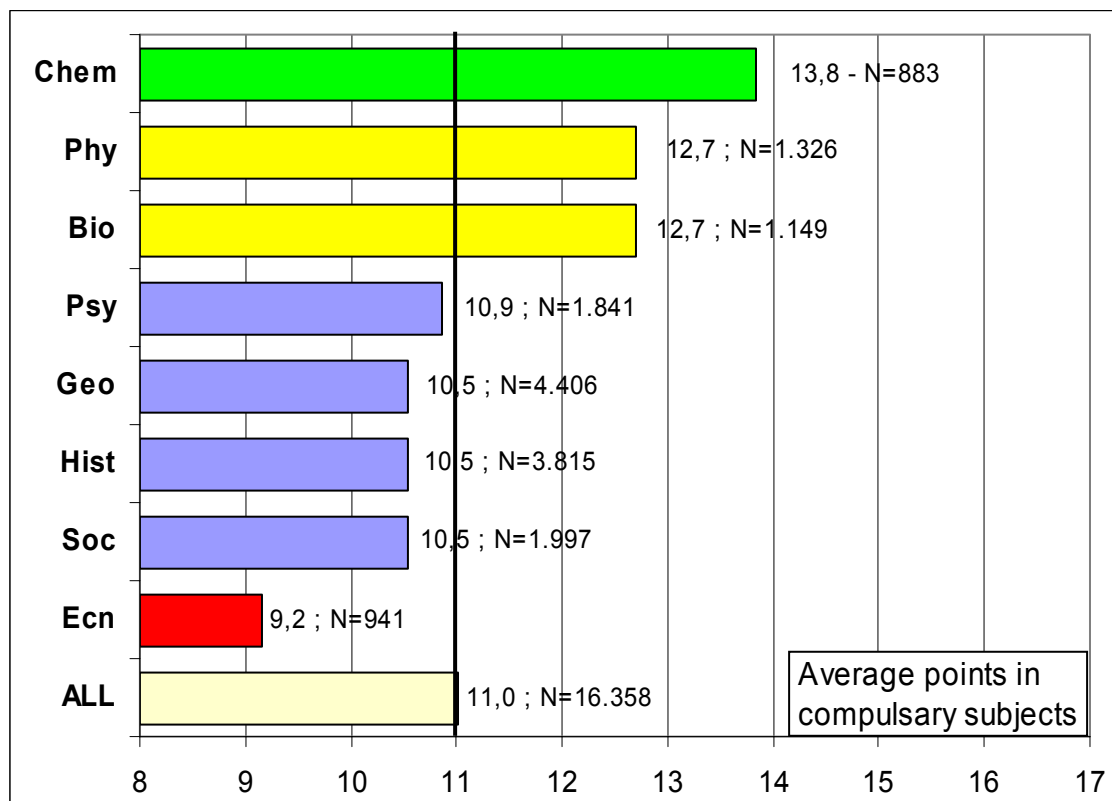


Figure 6. Candidates' average points in compulsory subjects in relation to the selection of the optional one (*Matura* 2004, Spring session)

Horizontal bars in Figure 6 show the average achievement of groups of students in the three compulsory subjects who selected a certain optional subject. Data are from the Spring session of *Matura* 2004. Data for other foreign languages, which candidates can also select in the optional part, and subjects with a smaller number of candidates, are not shown. Average sums of point grades in the three compulsory subjects for candidates with the following optional subjects are shown: Geography (N=4,406), History (N=3,815), Sociology (N=1,997), Psychology (N=1,841), Physics (N=1,326), Biology (N=1,149), Economy (N=941) and Chemistry (N=883). Subjects are listed according to the falling number of candidates who selected the subject for *Matura*. Number of candidates who select the so-called Social Sciences subjects is significantly higher (3.8 times) than the three Natural Sciences ones, i.e. Physics, Chemistry and Biology.

The average number of points for all Slovene students in the Spring session of *Matura* 2004 in the three compulsory subjects was 11.0. Figure 6 shows average points of groups of students in the three compulsory subjects who selected individual optional subjects. The eight groups of students (classified with regard to the optional subjects) are further classified in four groups according to the criterion of average achievement in the three compulsory subjects in *Matura*. The group with the highest average achievement in the three compulsory subjects is the group of students who selected Chemistry, with a small number of students taking the

exam on one hand but high average achievement of 13.8 on the other. The second group comprises the groups who selected Physics and Biology. The third group includes students with the following optional subjects: Psychology, Geography, History and Sociology; all subjects have a large number of candidates but very low average achievement, only slightly over 10.5. The fourth group includes students who selected Economy; the group is small and differs from the other groups also due to the fact that in general USSs candidates seldom select this subject. Thus, the average achievement is only 9.2.

Each group statistically significantly differs from other groups. Differences between the average number of points in the three compulsory subjects between the groups of candidates who selected Chemistry and Physics respectively is statistically significant, at the conservative 0.001 level of probability (t-test; $p=0.00000082$). Differences between the average number of points in the three compulsory subjects between the groups of candidates who selected Chemistry and Biology respectively is statistically significant, at the 0.001 level of probability (t-test; $p=0.00000157$). Differences between the average number of points in the three compulsory subjects in the group who selected Psychology and the three groups who selected Sociology, History and Geography at the conservative 0.001 level of probability, are statistically insignificant (t-test; $p=0.029$ (Soc), $p=0.014$ (Hist), $p=0.011$ (Geo)). So in this sense, we can deal with the students, who selected the four mentioned subjects, as one group only.

A step further can be made in the selection of subjects in *Matura*. As each candidate has to select two optional subjects, we can take a look also at the average number of points in the three compulsory subjects for (sub)groups of candidates who selected certain combinations of two optional subjects. There are two figures. Figure 7 shows the average achievement in the three compulsory subjects for several subgroups of candidates who in the combination of two optional subjects selected Physics. Figure 8 shows the average achievement in the three compulsory subjects for several subgroups of candidates who in the combination of two optional subjects selected Geography. As groups of candidates for Physics ($N=1,326$) and for Geography ($N=4,406$) are divided into smaller groups, the number of candidates in subgroups is accordingly smaller; nevertheless, it is evident that the average achievement in the three compulsory subjects is significantly different with regard to the aforementioned groups of candidates. With the exception of the combination Physics-Psychology, the gap between average points in the three compulsory subjects in the subgroup of Natural Sciences subjects and in the subgroup of Social Sciences subjects widens.

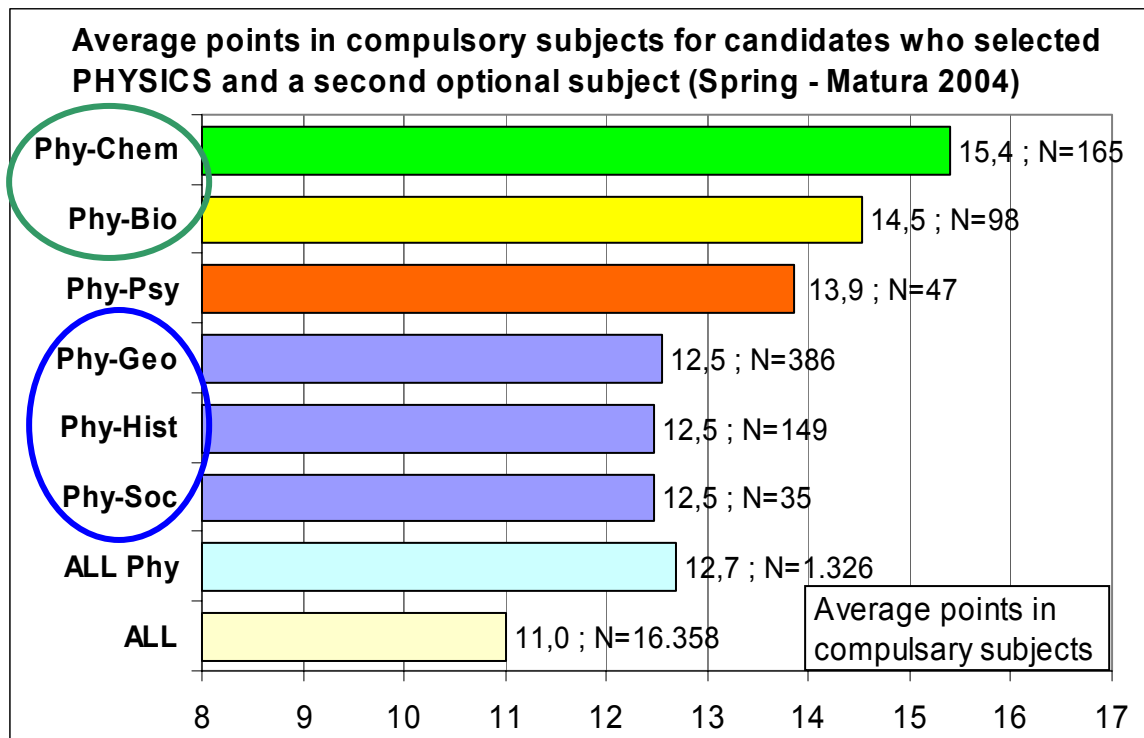


Figure 7. Average points in compulsory subjects for candidates who selected Physics and a second optional subject (*Matura* 2004, Spring session)

Figure 7 shows that the average number of points in the three compulsory subjects in the subgroup of candidates who selected the combinations Physics-Chemistry and Physics-Biology is significantly above the average, i.e. 15.4 and 14.5 respectively, with regard to all candidates who selected Physics (12.7). The combination Physics-Psychology is an exception. The average number of points in the three compulsory subjects in the subgroup of candidates who apart from Physics selected one of the Social Sciences subjects, either Geography, History or Sociology, is significantly below the average, i.e. 12.5 points, with regard to all candidates who selected Physics (12.7). Subgroups of candidates with two Natural Sciences subjects statistically significantly differ in average points in the three compulsory subjects from the subgroup of candidates who selected Physics and one Social Sciences subject, either Geography, History or Sociology (t-test; $p < 0.05$).

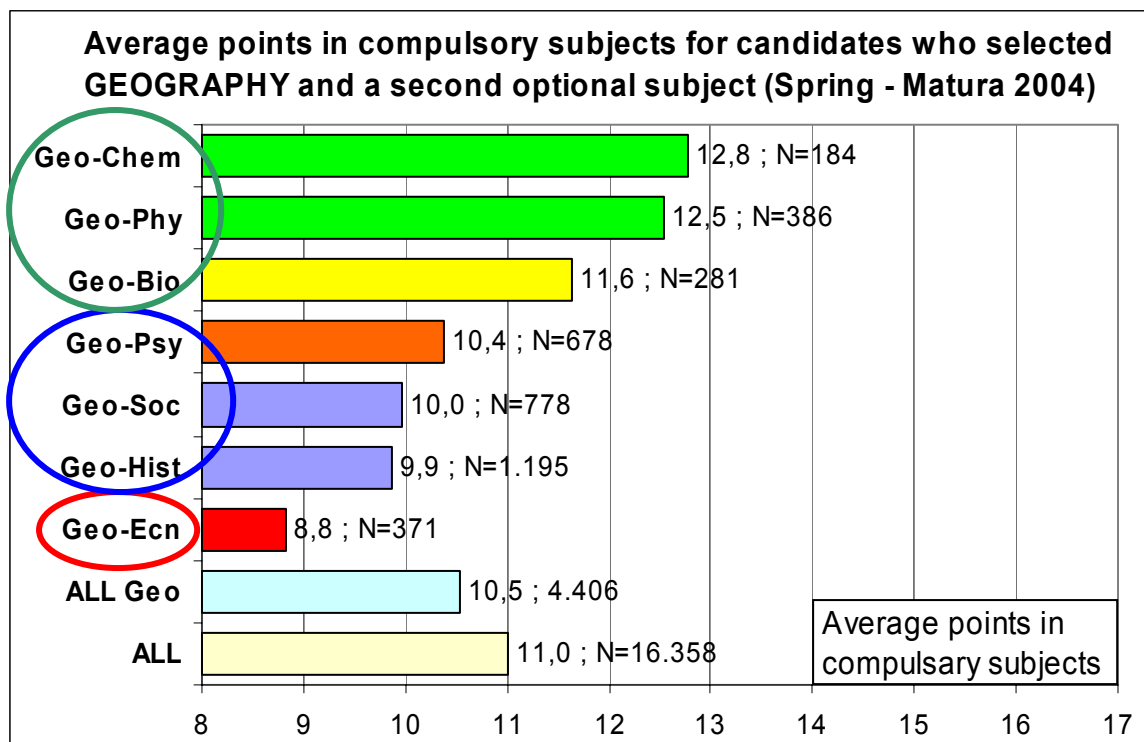


Figure 8. Average points in compulsory subjects for candidates who selected Geography and a second optional subject (*Matura 2004*, Spring session)

Figure 8 shows yet more distinctly that the average number of points in the three compulsory subjects of the subgroup of candidates who selected the combinations Geography-Chemistry and Geography-Physics or Geography-Biology is significantly above the average, i.e. 12.8 and 12.5 or 11.6 respectively, with regard to all candidates who selected Geography (10.5). The average number of points in the three compulsory subjects in the subgroup of candidates who apart from Geography selected another Social Sciences subject, either Psychology, Sociology or History, is significantly below the average, i.e. 10.4 and 10.0 or 9.9, with regard to all candidates who selected Geography (10.5). Candidates who apart from Geography selected Economy are again an independent group with the average number of points 8.8. Subgroups of candidates with Geography and a Natural Sciences subject statistically significantly differ in average points in the three compulsory subjects from the subgroup of candidates who selected two Social Sciences subjects (t-test; $p < 0.05$).

Natural Sciences subjects are (less frequently) selected by candidates with better achievement in the compulsory part, while Social Sciences subjects are selected in greater numbers by candidates with significantly lower achievement in the compulsory part. At the national level, these findings lead to a question of whether there is a pattern: some optional subjects (Chemistry, Physics and Biology as the complex of the three Natural Sciences subjects) are selected by candidates with better achievement in the compulsory common part (less frequently), while Social Sciences subjects (Geography, History, Sociology and Psychology) are selected en masse by candidates with significantly lower achievement in the compulsory common part. The selection of a Natural Sciences, a Social Sciences or a technical subject is statistically significantly linked with the achievement of candidates, who selected the subject, in the three compulsory subjects (Slovene, Maths and a foreign language). Do groups of subjects that candidates select present a kind of demarcation line between 'high', 'average' and 'low' achievers, at least on average? Such questions can be particularly relevant for the formulation of the national school policy in the light of national as well as European directives for increasing the interest of secondary-school students and students at universities

to study natural sciences and engineering which is one of the EU policies (Indicators on the quality of school education 2000). The findings are also important for the decisions that candidates make when entering university courses and that faculties make when admitting candidates who took different optional subjects in *Matura*.

Analyses at the national level are aimed at subject experts, members of national committees and all competent educational institutions in the country for evaluation studies and to help in decision making for improvement within the system of education.

Analysis: Trends in achievement in school

Each generation of students together with their parents and teachers as well as school management eagerly look forward to hearing about school grades and results of the external *Matura* exam. Due to the high-stakes role of *Matura* first reactions of individuals are always influenced by two facts: if a candidate passed *Matura* and if their achievement makes it possible for them to enrol on the desired course in tertiary education. School management too is, in the first place, interested in such basic data for the entire school but also in comparisons with the national average achievement. However, in order to establish how efficient a school is, it is not only the results of one session that count but achievement in a longer period, over several years. This is why longitudinal studies, i.e. detecting trends, are so important.

The ALAT Tool database comprises data for all schools for the last five years. Interesting trend analyses cover:

- overall achievement in school;
- average grades in subjects in school;
- average grades of candidates taught by the same teacher, etc.

As the national average can differ from year to year, the differences between years are shown as the average of achieved points, or the school grade (or class or teacher's grade), MINUS the national average of a comparable (sub)group.

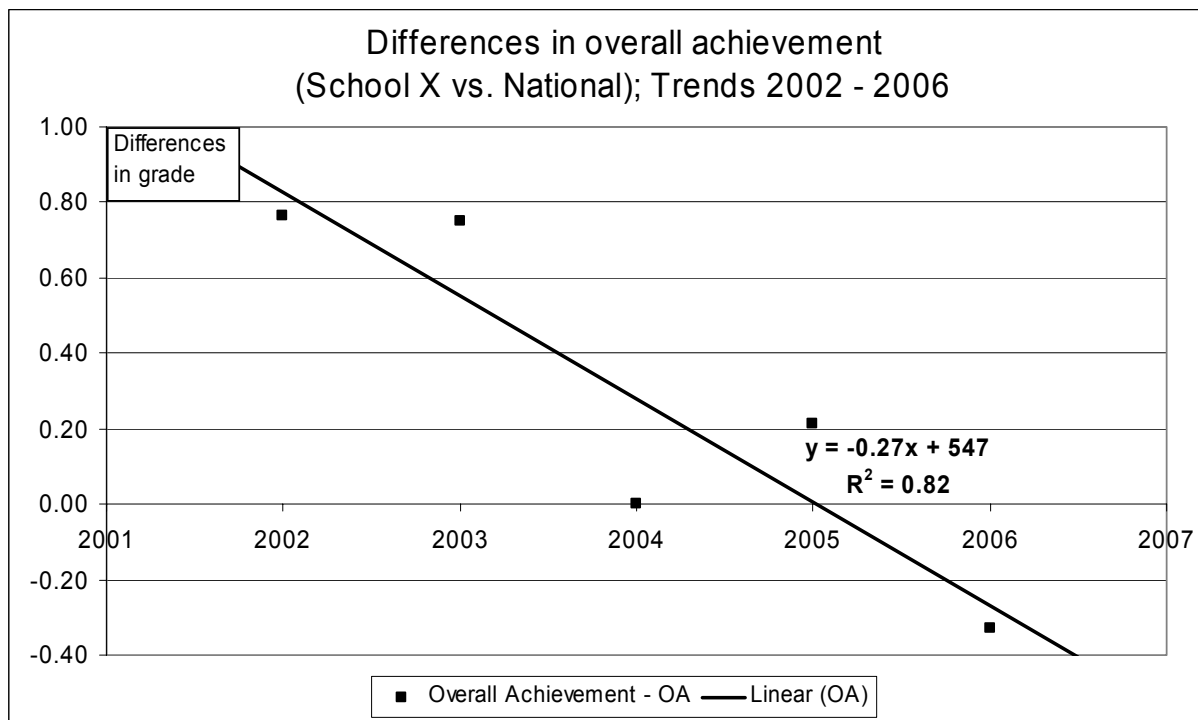


Figure 9. Differences in achievement between a school and the national average – trends

The regression line in Figure 9 shows the difference in achievement of a school and the Slovene average in five consecutive years of the Spring session of *Matura*, from 2002 until 2006. The case of the given school reveals how results, which were above the average, slid below the Slovene average.

The regression line in the picture is formulated as $y = m x + b$, where y represents the difference in the average overall achievement of candidates in one school and the average overall achievement of all candidates in general USSs in Slovenia; x represents the year, and m is the (negative) regression coefficient -0.27. The standard error of a slope is $SEM=0.088$ (Sanocki 2001). The explained proportion of variation by the year of the Spring session of *Matura* is high; the coefficient of determination is 0.82. The effect is real, not caused by chance alone, for $p < 0.05$ ($r=0.903 > 0.878$; degrees of freedom $df = 5 - 2 = 3$). Other variables account for only 18% of the difference in achievement of a school and the Slovene average.

Average point grades of candidates in the three compulsory subjects, i.e.: Slovene, Maths and English, are those that mostly contribute to overall achievement in the five-subject *Matura*. If a school detects a distinct (downward) trend in overall achievement, the ALAT Tool makes it possible to analyse trends in achievement in individual subjects. Figure 10 shows trends in achievement in the three compulsory subjects for the school with a distinct downward trend in overall achievement (shown in Figure 9).

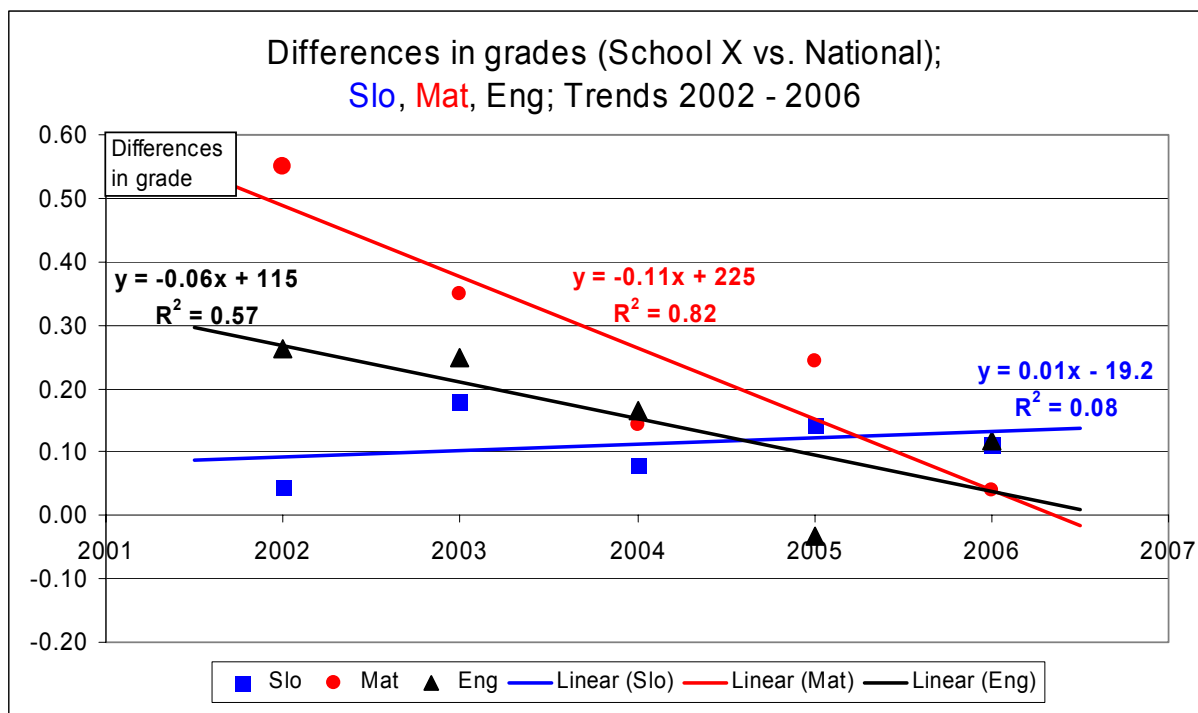


Figure 10. Differences in grades in Slovene, Maths and English – trends 2002–2006

The regression lines in Figure 10 show the difference in achievement of a school and the Slovene average in five consecutive years of the Spring session of *Matura*. The case of the given school reveals how results which were above the average in two subjects, English and Maths, slid below the Slovene average. The regression lines in the picture are formulated as $y = m x + b$; x represents the year, m is the regression coefficient – the slope – and b is the y intercept.

Value y in the first regression line (blue) represents the difference in the average point grade of candidates in one school in Slovene language (Slo), the mother tongue, and the average point grade of all candidates in general USSs in Slovenia in Slovene. The regression

coefficient is almost zero $m_{Slo}=0.01$. Average point grades in the school in Slovene had been stable in five consecutive years; the standard error of a slope is $SE_{m(Slo)}=0.048$. The explained proportion of variation by the year of the Spring session of *Matura* is low; the index of potential importance is $R^2_{Slo}=0.08$.

Value y in the second regression line (black) represents the difference in the average point grade of candidates in one school in English (Eng), a foreign language, and the average point grade of all candidates in general USSs in Slovenia in English. The regression coefficient is negative $m_{Eng}=-0.06$. Average point grades in the school in English had been decreasing in five consecutive years; the standard error of a slope is $SE_{m(Eng)}=0.051$. The explained proportion of variation by the year of *Matura* is $R^2_{Eng}=0.57$.

Value y in the third regression line (red) represents the difference in the average point grade of candidates in one school in Maths (Mat) and the average point grade of all candidates in general USSs in Slovenia in Maths. The regression coefficient is negative $m_{Mat}=-0.11$. Average point grades in the school in Maths had been decreasing in five consecutive years; the standard error of a slope is $SE_{m(Mat)}=0.035$. The explained proportion of variation by the year of *Matura* is high; the index of potential importance is $R^2_{Mat}=0.82$; it explains almost all the variation in the difference in the average point grade of candidates in the school in Maths. The correlation is significant for $p<0.05$ ($r=0.907>0.878$; $df = 3$) (Sanocki, 2001).

Analysis: Differences in average grades between subjects in a school

At the school level, the ALAT Tool allows schools to compare their achievement, i.e. average grades in *Matura* with average grades of a comparable group at the national level. The histogram in Figure 11 shows five subjects, the three compulsory ones (Slovene, Maths and English) and two from the wide selection of optional ones (Chemistry and History). Each column shows the deviation of the average grade in a subject in a school from the average grade of all candidates who took an exam in this subject in *Matura* in the same educational programme in Slovenia.

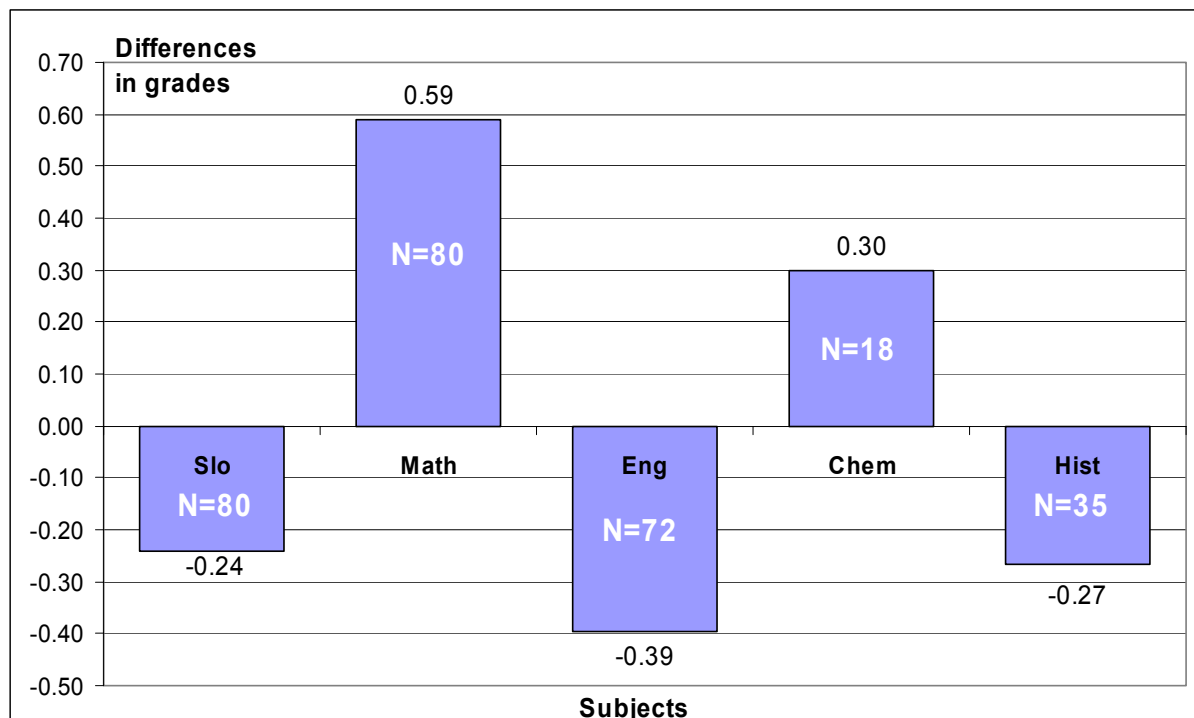


Figure 11. Differences in average grades (an USS vs National)

The histogram shows results for an USS with five classes in the Spring session of *Matura* 2006. The marking scale has five grades and the differences in grades from 0.25 of the grade to 0.60 of the grade are significant. The shown case is authentic: in Maths and Chemistry, the school is above the national average, while in Slovene, English and History it is below it. Reezigt, Guldemond and Creemers (1999) found that although there were some effects on students' achievement of classroom and school level factors these were by no means stable across subject areas. The aim of such monitoring of the results of *Matura* and comparisons is to open a discussion among experts and to interpret possible causes to lead to a decision for a change and improvement in further work. Decisions for further work leading to improvement and greater efficiency, however, cannot be general and the same for all: for the entire school, for all subjects, in every teacher's classroom and for different classrooms of the school. On the contrary, the adopted measures have to be specific and have to vary from subject to subject and from teacher to teacher but have to always be based on analyses and self-evaluation of teachers in the school.

Analysis: Variability between schools and between classes within a school

When in school-based assessment and in external assessment such as Slovene *Matura* achievement is reported on, the professional public, the press and parents are usually given information on the average achievement of students at the school level. Students in individual schools, however, are nested in their basic units, i.e. the classroom. Academic studies and practitioners too find that differences in achievement between classes within a school are equally important, if not even more important, than differences between schools. Since data on students' school grades and their achievement in *Matura* are gathered also at the classroom level within schools and schools are familiar with the fact which teacher taught which student, the ALAT Tool allows for different data analyses of classrooms and/or teachers.

Figure 12 shows variability in schools and classes of average grades in Slovene (i.e. the mother tongue) which is compulsory for all students taking *Matura*. The grades are represented on a scale from 1 to 5, where 1 is Fail and 5 is the highest grade (Excellent). Each horizontal bar represents one school, i.e. one general USS in Slovenia, in the Spring session of *Matura* 2006. The vertical line inside the bar marks the average *Matura* grade in the subject in an individual school. Schools have anything from one and up to nine classes, the average being 4.6 classes a school. Students' achievement in classes within the same school can vary considerably. The left-hand edge of each bar marks the average grade of students in the class, whose result was the lowest, while the right-hand edge of the bar marks the average grade in the subject in the class, whose result was the highest.

The highest position in Figure 12 is taken by the bar, which has the value of the lowest average grade in the school in Slovene (2.49) on the left-hand side, and the value of the highest average grade of USSs in the subject (4.16) on the right-hand side. Second in the graph is the bar with the lowest average grade of the class (2.25) on the left-hand side, and the highest average grade of the class in Slovene (4.38) in all general USSs in Slovenia in the Spring session of *Matura* 2006 on the right-hand side. It is expected that variability in average grades between classes in the entire country is greater than variability between schools.

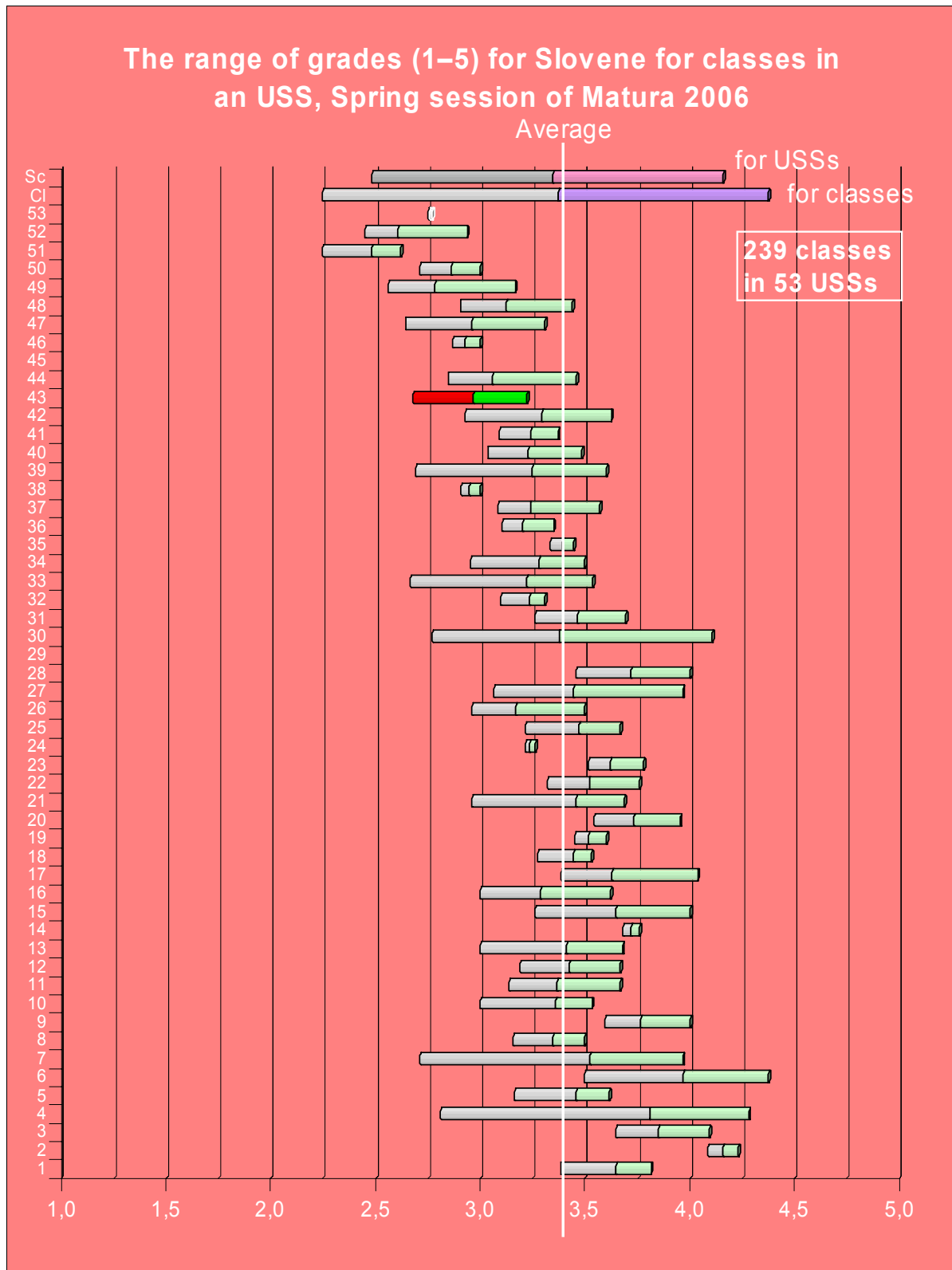


Figure 12. Variability between schools and between classes within a school for Slovene

A school may have many classes so average grades in one subject in *Matura* (e.g. Slovene) reveal no statistically significant differences (t-test; $p < 0.05$). Data for the USS with a serial number 1 with nine classes show that the average grade in the school is 3.65; the highest average grade of a class is 3.81, and the lowest average grade of a class is 3.38. In this case, the differences in average grades between the nine classes are statistically insignificant (t-test;

$p < 0.05$), except between the two classes with the highest averages and the one with the lowest average in school (t-test; $p = 0.024$ and $p = 0.013$).

Data for the USS with a serial number 4 with eight classes shows a different picture. The average grade in Slovene in the school is 3.81 and is higher than in the USS no 1 (as the average grade of the USS no 4 is the same as the average grade of the best class of the USS no 1). The highest average grade of a class is 4.28, which is the second best result between all classes in all Slovene USSs (ranking 2nd or in 99th percentile). On the other hand, the same school has a class with the lowest average grade of 2.81, which is well below Slovene average (ranking 221st from 239 classes or in 7th percentile). The differences in average grades between the eight classes of the school (i.e. each class compared to all the others) are statistically significant (t-test, $p < 0.05$) in more than one half of the cases, i.e. in 15 combinations (pairs) out of the possible 28.

Let us take a look at the USS with a serial number 43 with six classes and below-average achievement in Slovene. The average grade of the school is 2.97, the highest average grade of a class is 3.22, and the lowest average grade of a class is 2.68. The differences in average grades between classes are statistically insignificant (t-test, $p < 0.05$), except between the two classes with the highest and the one with the lowest averages in school (t-test; $p = 0.043$ and $p = 0.009$).

So what average grades does the same USS (serial number 43) get in the same exam session in another compulsory subject, in Maths – Figure 13? The differences in average grades between classes are statistically insignificant (t-test, $p < 0.05$), except in the class with the highest average and three other classes in school (t-test; $p = 0.001$ and $p = 0.001$ and $p = 0.002$).

The USS with a serial number 43 with six classes has average achievement in Maths. The average grade in the school is 3.74 (and Slovene average is 3.73); the lowest average grade of a class is the below-average 3.36, while the highest average grade of a class is 4.31, which is high even for USSs with the highest average grades in Maths in *Matura* (ranking 20th from 239 classes or 91st percentile).

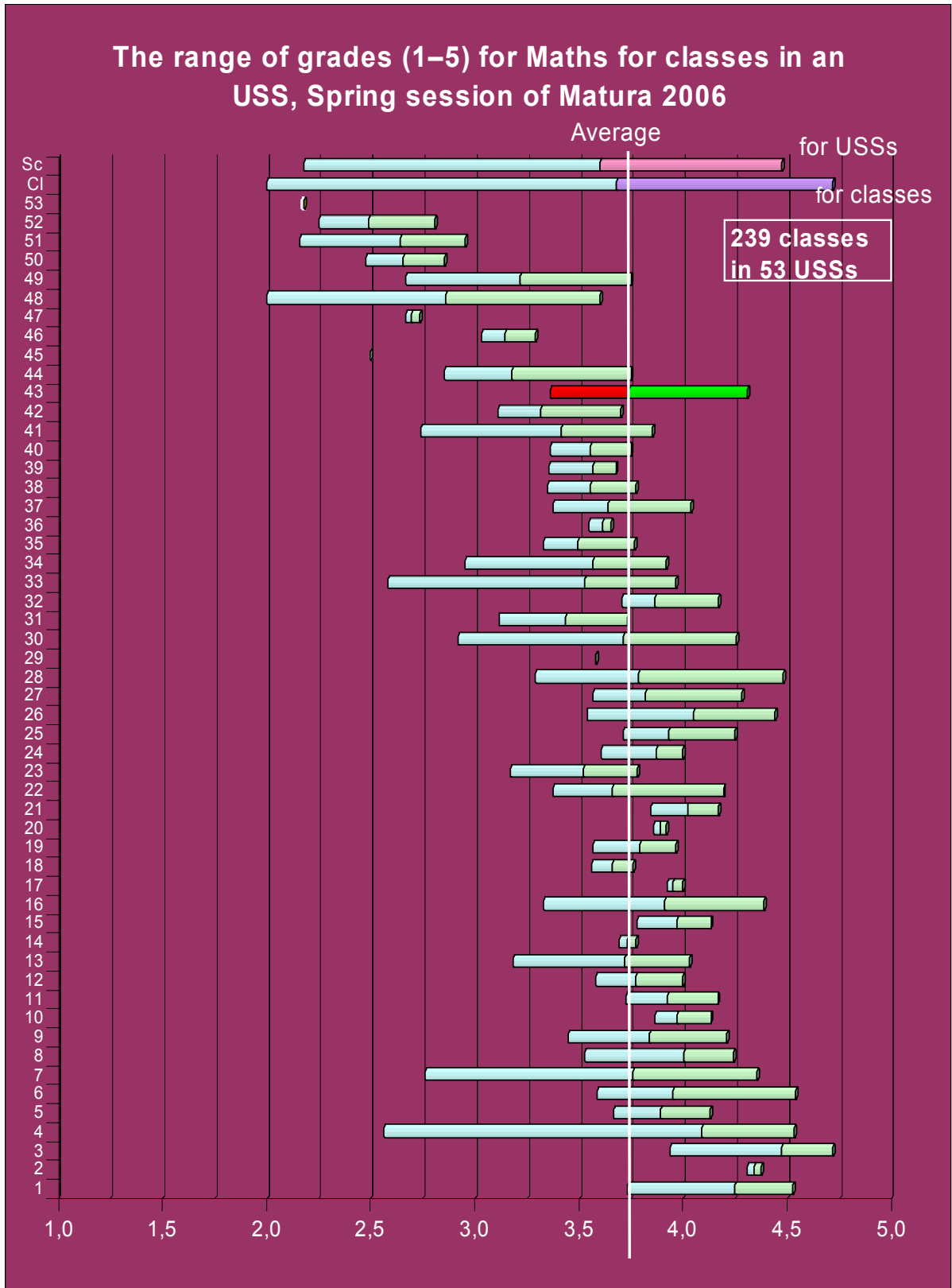


Figure 13. Variability between schools and between classes within a school for Maths

Although variability in achievement between different schools is expected to some extent because students enrol in secondary school programmes also according to achievement in primary education, significant differences in achievement in different classes within individual schools have been established too, despite the central tenet that classes in schools

are not composed in accordance with students' abilities. Comparison of achievement of all pairs of classes within schools reveals that there can be frequent significant differences between average grade in one subject, but not in another.

Analysis: Average classroom grades

Schools are made up of classrooms and students and in addressing areas of effectiveness it is often wiser to look further into the school rather than at the school (Webster, & Fisher, 2000). Two frequency distributions of average grades are shown in Figure 14, for Slovene (magenta) and Maths (blue), which represent the number of classes where students got the average grade with the intervals of 0.1 on the marking scale from 2 to 5 for each subject. Average grades have been calculated for classes (236 classes in 50 USSs) for one exam session (i.e. the Spring session) of *Matura* 2006. Within each of the distributions, the six classes of the school are shown on the abscissa. There are differences between subjects. The average grade in one school in Slovene was below the average, 2.97, while the national average was 3.40. The average grade in the same school in Maths was the average 3.74, as the national average was 3.73. One would expect average grades in different classes of the same school to be approximately the same, that one school is 'hidden' within one column in the distribution or maybe in two adjacent ones. In this instance this is not the case; the distribution of average grades in a subject (Maths or Slovene) between classes of the same school is wide. Variability of achievement between classes of one school (six classes, i.e. A, B, C, D, E, F) is great too. Average grades in classes of the school are distributed between 2.7 and 3.3 in Slovene and 3.4 and 4.4 in Maths.

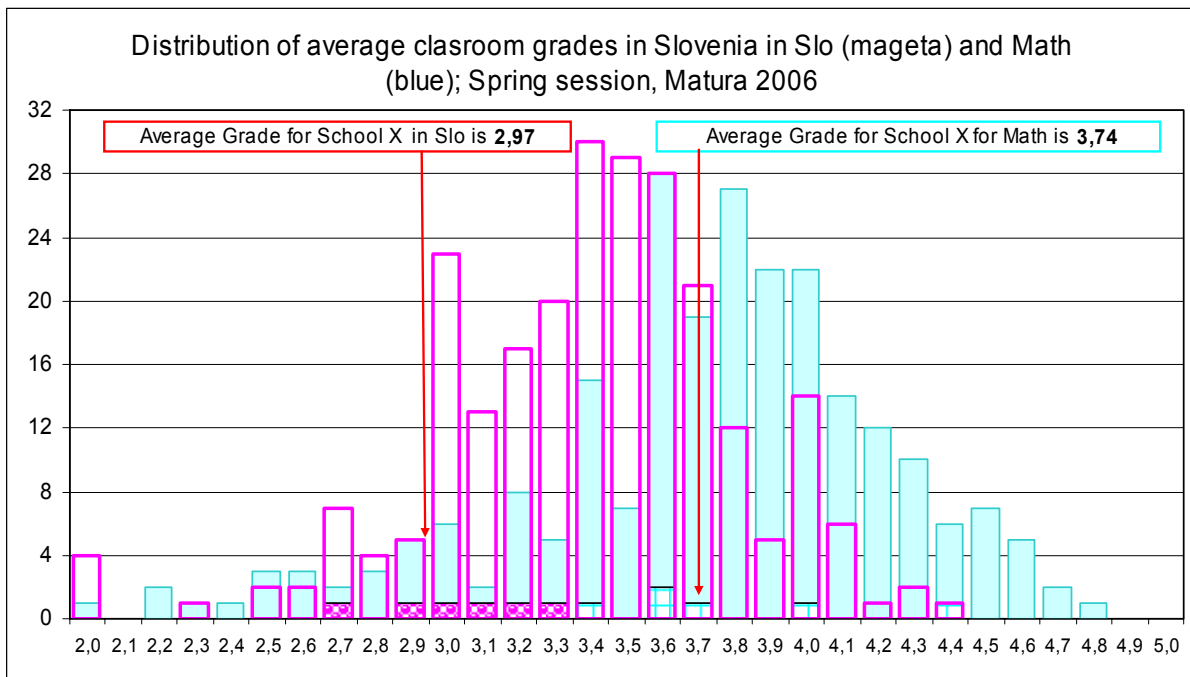


Figure 14. Distribution of average classroom grades in Slovenia in Slovene (magenta) and Maths (blue)

There are great differences between the subjects in the school. Average grades of all classes in the school in Slovene are below the national average in this exam session, while in Maths there are some classes below the national average and some above the average, one even well above it with the average grade of 4.4.

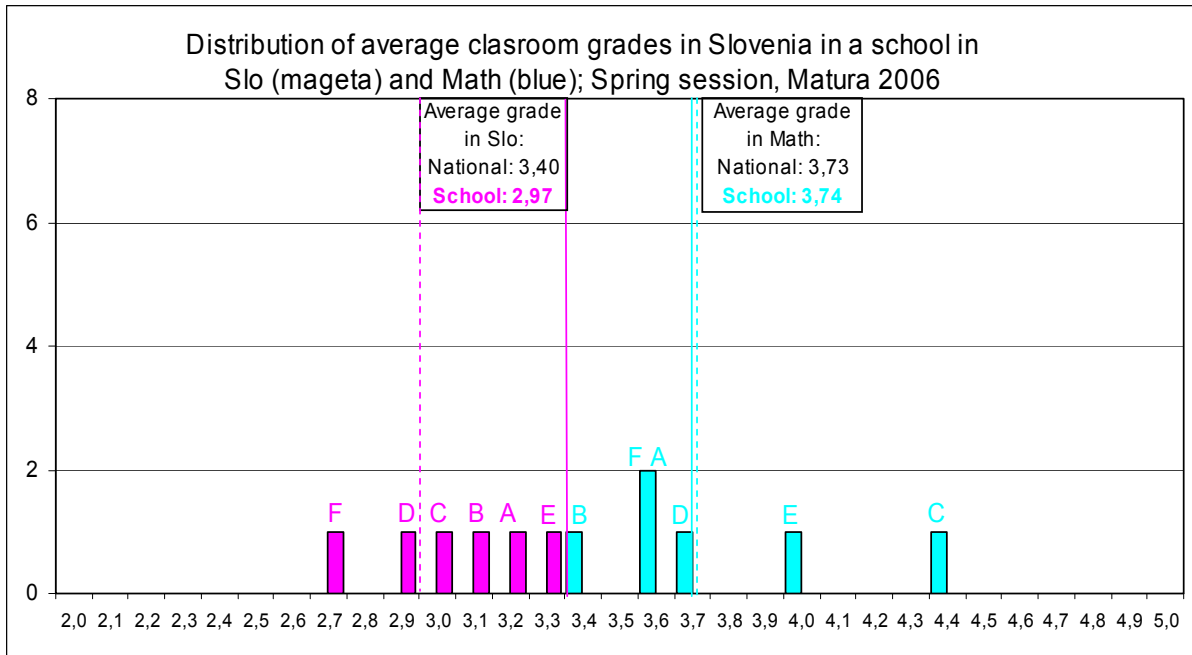


Figure 15. Distribution of average classroom grades in an USS in Slovene (magenta) and Maths (blue)

It is interesting to take a closer look at individual classes in the school. Are the classes in which students have above-average achievement in Maths better achievers in Slovene too, at least within the same school? Is the sequence of classes according to average grades in *Matura* for these two subjects the same? In the shown case, the following is true of students' achievement in different classes in the same school:

- students' achievement in different classes is below the average in Slovene and average in Maths;
- the distribution of classes according to average grades in subjects is very dispersed, from low to very high grades; and
- students who get high (low) average grades in one subject do not necessarily get the same grades in another.

The sequence of classes in the school according to average grades in Maths is different than that in Slovene.

Rank order in Slovene	Classes	Rank order in Maths
1	4E	2
2	4A	4
3	4B	6
4	4C	1
5	4D	3
6	4F	5

Table 1. Different rank orders of classes in two compulsory subjects: Slovene and Maths

An analysis of school grades at the end of schooling (i.e. prior to *Matura*) was performed for the school and classes in question. Classes' ranks according to average school grades at the end of schooling were different from ranks according to grades achieved in *Matura*. This shows that in this school the classes were not ranked according to the intensity of teaching the

mother tongue (i.e. Slovene) or Maths. If a school or its class has high (low) average in one subject it does not necessarily have high (low) achievement in another as well.

A variability analysis of the average number of achieved per cent in the external part of *Matura* 2006 for Maths and Slovene for all classes in Slovene USSs in the Spring session of *Matura* 2006 has been performed. For the first time, students sat the exams in 50 USSs with a different number of classes, from two and up to nine, totalling 236. There were 545 possible combinations of pairs of classes within individual schools. It turned out that in 161 cases for Maths (and 126 for Slovene), or 30% (and 23% for Slovene), there was a statistically significant difference between average achievement in Maths (and Slovene) between pairs of classes within the same school, for $p < 0.05$ (t-test).

				School X					School Y					School Z					School W						
Class	Nu	%Eks	STD	4a	4b	4c	4e	4f	4g	4h	4a	4b	4c	4d	4e	4f	4a	4b	4c	4d	4e	4f	4g	4.a	4.b
4a	23	54,9	12,5		0,871	0,355	0,350	0,012	0,489	0,595															
4b	26	53,3	13,5			0,181	0,165	0,003	0,808	0,759															
4c	24	58,3	12,6				0,937	0,121	0,098	0,317															
4e	27	58	10,9					0,065	0,080	0,327															
4f	24	63,3	9,18						0,001	0,101															
4g	27	52,4	12,2							0,854															
4h	7	51,1	16,3																						
4a	10	58,6	13								0,946	0,039	0,783	0,271	0,668										
4b	11	58,3	8,13									0,002	0,627	0,095	0,460										
4c	13	68,8	5,31										0,004	0,064	0,001										
4d	11	59,9	7,41											0,205	0,820										
4e	18	63,9	8,86												0,240										
4f	25	60,6	9,16																						
4a	31	65,9	8,31														0,002	0,001	0,002	0,000	0,074	0,381			
4b	31	71,6	5,39															0,682	0,818	0,040	0,170	0,000			
4c	32	72,3	6,26																0,878	0,140	0,098	0,000			
4d	32	72	6,72																	0,115	0,144	0,000			
4e	32	74,3	4,59																			0,002	0,000		
4f	33	69,5	7,1																				0,008		
4g	32	64,1	8,64																						
4.a	24	70,3	5,23																						0,275
4.b	26	68,5	6,15																						

Table 2. Significant differences in average achievement between classes (yellow) in four USSs, X, Y, Z and W (t-test; if $p < 0.05$)

A variability analysis of the average number of achieved per cent in the external part of *Matura* 2006 for Maths and Slovene for USSs has been performed, too. There were $(50 \times 49) : 2 = 1,225$ possible combinations of pairs of USSs. It turned out that in 824 cases for Maths (and 785 for Slovene), or 67% (and 64% for Slovene), there was a statistically significant difference between average achievement in Maths (and Slovene) between the USSs; for $p < 0.05$ (t-test). The main reason for the differences between USSs was that there were differences in achievement at the end of primary (lower secondary) schools between students enrolled. It is evident too, however, that there are significant differences also between classes within individual schools despite the practice that classes in secondary schools are not composed in accordance with students' achievement in primary education.

Analysis: Different teachers' grades

Fullan (2000) said that

"one of the strategies in redefining accountability is that teachers and heads become more "assessment literate" and assessment becomes more transparent."

The basic organisation unit in education is a group of students in a class headed by a teacher. This is where differences appear. Analyses of classes and groups within schools can be performed according to grades before *Matura* or according to points or grades achieved in *Matura*. The ALAT Tool makes it possible to make comparisons between school grades and

grades in external exams such as *Matura*. This makes external assessment increasingly more transparent. The ALAT Tool allows for the identification of variability at the classroom level as well and is a good starting point for self-evaluation of teachers' work in classes within a school. The use of the ALAT Tool should make teachers and head teachers more "assessment literate".

Analyses of comparisons of school grades and grades achieved by students in an external exam such as *Matura* are of special importance. The case (Figure 16) illustrates students' achievement with three teachers in a school where Geography was taught in eight classes. The subject was selected for *Matura* by 34 students (red) from three classes (4A, 4B and 4C) taught by the first teacher; 31 students (green) from three classes (4Č, 4D and 4E) taught by the second teacher; and as many as 42 students (yellow) from two classes only (4F and 4G) taught by the third teacher. Students in all these classes had the same number of Geography lessons in the four year-period of USE and consequently the differences in overall achievement at the end of USE between them were not great. The average overall achievement in Year 4 was 3.64 (STD=0.87) in the red group of classes (4A, 4B and 4C) with 88 students; 3.86 (STD=0.88) in the green group of classes (4Č, 4D and 4E) with 94 students; and 3.80 (STD=0.75) in the yellow group of classes (4F and 4G) with 59 students.

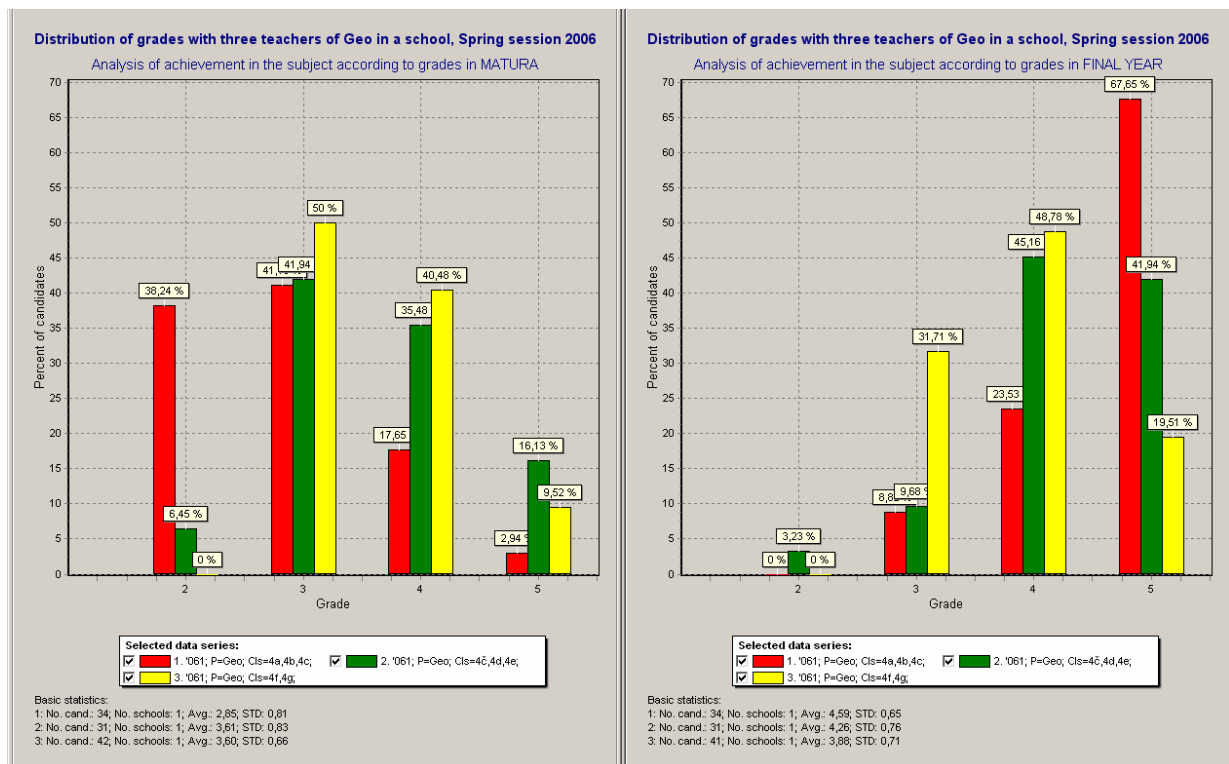


Figure 16. Distribution of grades with three Geography teachers in the same school (*Matura* 2006, Spring session): Analysis of achievement in the subject according to grades in *Matura* (left) and according to grades in Year 4 (right)

The left-hand diagram shows the distribution of grades (1–5) in Geography for the three groups of students (taught by different teachers of the same school) in the Spring session of *Matura* 2006. The right-hand diagram shows the distribution of school grades (1–5) at the end of schooling in Geography for the same groups of students; this is how their knowledge was graded by their teachers prior to *Matura*. The grades in *Matura* are lower from school (i.e. teachers') grades in all three groups. The average grade in *Matura*, 2.85, is the lowest in the first group (school grade: 4.59); the average grade in *Matura* in the second group is 3.61 (school grade: 4.26); and in the third group the difference between the grades (in *Matura* 3.60

and in school 3.88) is the least great. At the national level, no differences have been perceived between the two average grades, in *Matura* (3.49) and in classroom (3.48).

The differences in average grades in *Matura* between two groups of students – the green one and the yellow one – are not great, although the teacher's grade of the students from the green group at the end of schooling was much higher. A very low average grade in Geography in *Matura* was obtained by the students from the red group, 2.85, which is statistically significantly lower (t-test; $p < 0.05$) than the achievement of students in the school in the other two groups, the green one, 3.61 ($p = 0.0004$), and the yellow one, 3.60 ($p = 0.00005$). Data on teachers' grades at the end of schooling, i.e. prior to taking *Matura*, show, however, that variability of achievement between groups and teachers in the school is even more explicit. Students' knowledge of geography in the red group was awarded the highest average grade by their teacher, 4.59, which is higher than in the green group (4.26; $p = 0.066$), and the difference is even greater and statistically significant when compared to the yellow group with the average of 3.88 ($p = 0.00002$).

The bars (on the left) in both diagrams in Figure 16 show the distribution of grades of one Geography teacher. The teacher in question did not give Grade 2 (Satisfactory, i.e. Pass) to any of the students, all grades are higher: 8.8% got Grade 3 (Good), 23.5% got Grade 4 (Very Good) and 67.7% got Grade 5 (Excellent) – the right-hand diagram. More than two thirds of the students (23) got the highest grade, Excellent (5). In the external exam, however, the distribution of grades is 'inversed': more than one third of the students (38.2%) got Satisfactory (2), and only one (1) student got Excellent (5) in *Matura*, i.e. 2.9% – the left-hand diagram. Identification of such variability which cannot be a matter of coincidence is a good starting point for self-evaluation of teachers' work in classes within a school.

Analysis: Selecting a subject

Analyses are important for the national school policy, teachers' work in classes and in the school in the area of external exams such as *Matura* where candidates have to sit exams in several (five) subjects, some (two) of which they can select. This invites a string of questions such as:

- Which subjects do students select and which not? Are there any significant differences from school to school, from class to class? Do students select the optional subjects and thus indirectly their university course based on the popularity of the subject or its teacher in the school?
- What proportions of students select certain subjects at the national, school or classroom level or with a certain teacher? We could say that when students select subjects for *Matura*, they in a way 'give their vote' for subjects (at the national level) and for teachers of these subjects within individual schools. Normally, teachers wish that students select 'their' subject for *Matura*. If a high proportion of a teacher's students select a certain subject and have outstanding achievement in *Matura* (in comparison with a comparable group), then this is definitely a good indicator that a teacher's work is efficient and of high quality.
- What is the structure of students who select certain subjects? Are Natural Sciences subjects selected mostly by boys or also frequently by girls? Are individual Social Sciences subjects selected by students with high and low achievement or, on average, by students whose overall achievement is low?
- What is the achievement of students in the subjects they select? This is also linked with the proportion of students who select a subject. There is an important difference between the two: (a) if a certain subject is selected by a small number of highly capable or efficient students and they have solid achievement in *Matura*, or (b) if a teacher of a subject

manages to 'attract' a large number of students whose achievement is average as well as below and above the average and they achieve good results in *Matura*.

In Slovenia, students select the tier (either Foundation or Higher) at which to take the exam in two of the compulsory subjects, i.e. Maths and the foreign language (any of the following six: English, German, Italian, Spanish, French, or Russian). Therefore, the mentioned analyses, i.e. of the tier that students select, are important for these subjects as well.

In Slovene general USSs with a large number of candidates and several classes in Year 4, a student can select two optional subjects from among more than ten for *Matura*. On one of the schools, candidates from six different classes (4A, 4B, /.../ 4E) had the same number of chemistry lessons before they made the decision on which optional subject they would select. In 2006, candidates' decisions on Chemistry as one of the two optional subjects in the Spring session of *Matura* were different. Figure 17 shows the proportions of students from individual classes within the school who selected Chemistry as an optional subject in *Matura*. At the national level 16% of students in general USSs selected Chemistry in 2006. The differences between classes within the mentioned school are substantial, from 36.7% in Class 4A to Class 4B where none of the students selected Chemistry for *Matura*.

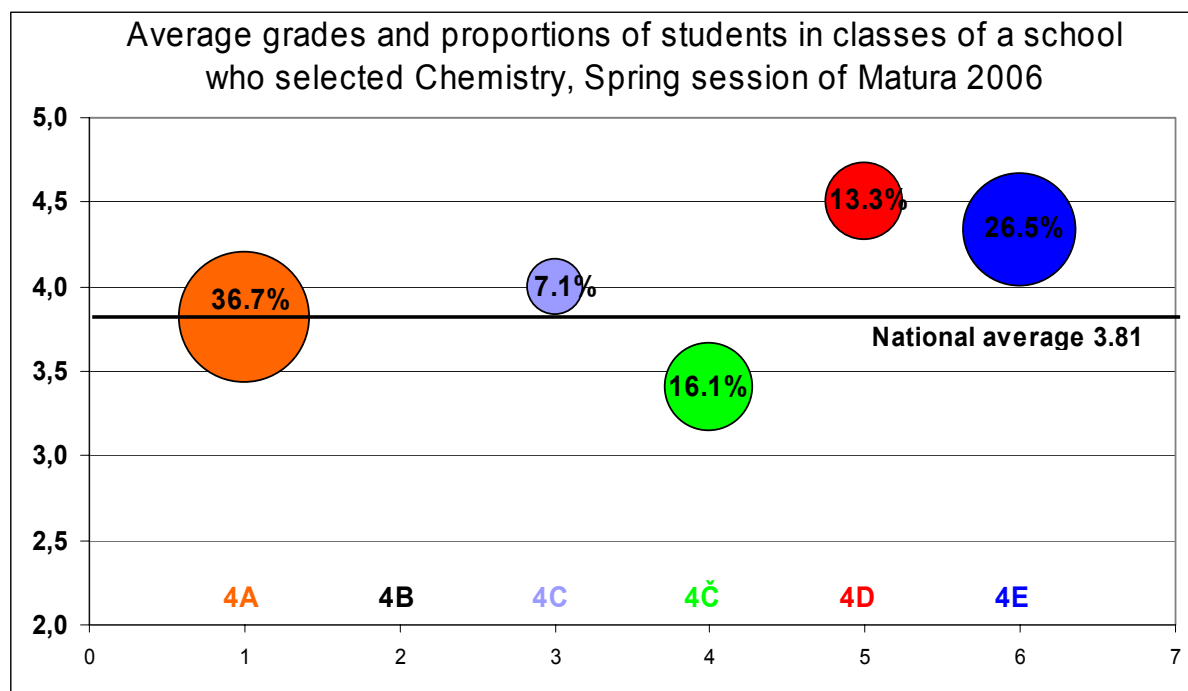


Figure 17. Average grades and proportions of students in classes of an USS who selected Chemistry (*Matura* 2006, Spring session)

Figure 17 also shows average grades that groups of students in individual classes obtained in Chemistry in *Matura* 2006. The class with the highest proportion of students who selected Chemistry obtained the average grade of 3.82, which is on the level of the national average (3.81 on the marking scale from 1 to 5). Students from two classes of the school (4C with 7.1% and 4D with 13.3%) on average obtained above-average grades in Chemistry, i.e. 4.0 in 4C and 4.5 in 4D, although the proportion of students in these two classes was lower than the national average of 16.0%. More than one fourth of students of Class 4E (26.5%) selected Chemistry in *Matura* and on average they too, this time with a high proportion, obtained a high grade of 4.3. The lowest average grade, below the national average, was obtained by the students from Class 4Č, where the proportion of students who selected Chemistry, i.e. 16.1%,

was in the national average. Due to the small number of students who within a class in a year select any given subject, it is not adequate to perform statistical analyses. In the described case, Chemistry was selected by only two students from Class 4C, while the largest number of students who selected Chemistry was in Class 4A, eleven.

Such analyses are intended for schools and their teachers as they are the only people who can be familiar with the causes of such variability. The nature of these causes can sometimes be easily explained and there is no need for a change; but on the other hand, such analyses, especially if monitoring is held from year to year, can lead to a better understanding of the causes and consequently to improvement as well.

Conclusions

With variability analysis at the national, school and classroom levels in USSs we want to share our experience in policy principles for improving learning in secondary classrooms and schools. For evidence-based decision making it is important to ensure that classroom, school and system level evaluations are linked and are used formatively to shape improvements at each level. Stronger bridges between research, policy and practice should be built, summative and formative assessment approaches should be aligned, the focus on teaching and learning should be kept and innovation should be encouraged (*Formative assessment*, 2005). We believe that using data and the tool for an effective change in schools does not come about by legislation nor by sanctions and punishment for failure. We trust that effective change will occur as schools, over time, experience the effects their planning, implementing, assessing and interventions have had in improved learning outcomes (Sliwka, Fushell, Gauthier, & Johnson, 2005). Key messages communicated to the schools must be that the "*power of the data are not so much in the numbers but in the discussions related to those numbers*" and that "*the data do not so much provide the answers, but they can inform the good questions for further investigations*". The established significant discrepancies at different levels of the school system reveal the source of potential problems and thus bring up relevant questions on the effectiveness as well as form the basis for deciding on improvements.

Hargreaves and Fullan (1998) recommend that teachers must become more "assessment literate". Assessment literacy refers to the capacity of teachers – alone and together – to examine and accurately understand student work and performance data and correspondingly to develop classroom and school plans to alter conditions necessary to achieve better results. The role of the learning community in education (a group of teachers of the same subject, class, year, the head teacher,...) is to extract relevant information from results of analyses of achieved curriculum for more effective teaching in the classroom and for learning support.

Member States in the European Union agreed that through evaluation and steering, schools can measure themselves against other comparable institutions. All countries are seeking the best way in which to report school performance, and to this end they are using internal or external evaluation or a system which combines both (*Indicators on the quality of school education* 2000). Accountability may be regarded as a two-stage process involving first, the identification of the performance of the education system in relations to its goals, as defined at any one time, and second, the response by educational institutions brought about through the mechanisms of system control in response to any perceived shortfall between performance and goals (Broadfoot 1996). Although conceptually distinct, these two stages should become simultaneous in practice.

Data gathered in both summative and formative processes are vital to understanding whether individual teachers, schools, and systems are meeting the goals. Connected data at different levels (national, school, classroom, student) should support each other in order to improve teachers' instruction and students' achievement – there should be consistency of effectiveness, characteristic within and between levels (Creemers 1994). Policies that link range of well-

aligned and thoughtfully developed assessment at the classroom, school and system levels will provide stakeholders with a better idea as to what extent they are achieving objectives. Policy and school leaders and teachers will have a sound basis on which to make improvements, and will broaden teaching as well as policy repertoires (*Formative assessment, -2005*).

The models of decision making, feedback regulation systems, effectiveness and improvement have similarities with curriculum, teaching and learning processes. Deming (1986) proposed that processes should be placed in a continuous feedback loop to analyse and identify sources of variation at every step in a process that cause outcomes to deviate from the expected and change the parts of the process that need improvements. The feedback data should be assessed to determine the causes of variation at each step in a process, to identify significant problems, and then focus attention on improvement. Deming created a diagram to illustrate this continuous process, commonly known as the PDCA cycle for Plan, Do, Check, Act.

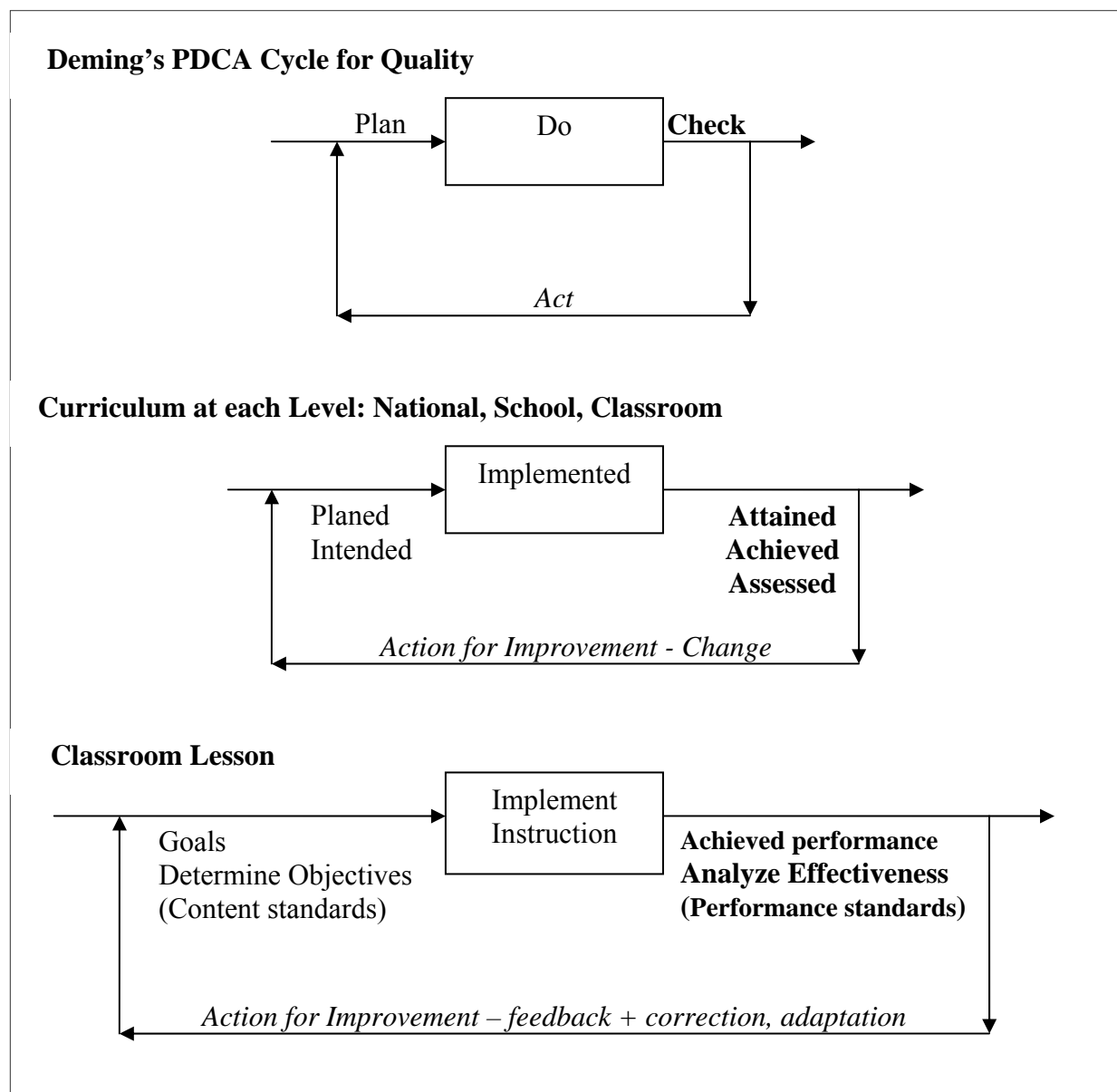


Figure 18. Circular causality of feedback loops; planning, establishing and assuring the quality of education at all levels: national, school, classroom and individual

It is similar in education. A curriculum can be planned at the national level, but undoubtedly at the school level and every teacher makes plans at the classroom level as well, the so-called intended curriculum. The central part in the process belongs to its implementation in practice, the so-called implemented curriculum. It is important to be familiar with the initial knowledge, motivation and the expectations of students and teachers, with what is in reality taught in schools and classrooms, with how teachers teach, with how and how much students learn, with the vast opportunities for teaching and learning. In all processes it is important to continuously assess (check) the outcomes, the so-called attained or achieved curriculum (Mullis, Martin, Gonzales, & Chrostowski, 2004). Analyses of what had been achieved can, or must, provide feedback information and correction or adaptation for evidence-based decision making for changes – continuous actions for improvement. Educational policy steering, changes in intended and implemented curriculum must be grounded in findings deriving from the attained curriculum.

Educational improvements begin with a clear idea of what students are expected to learn. Classroom lesson objectives (content standards) are statements that specify what teachers should teach and what students should learn. Standards specify goals or expectations and there may be many ways in instruction to implement the ends identified in the objectives. Effective standards must be specific enough to assess (check) when the standards have been achieved. Content standards specify 'what' while performance standards specify 'how well' – they answer the question: *"How good is good enough?"* (Linn, & Gronlund, 2000). Assessment ought to provide information on the extent to which standards have been achieved, while analyses ought to offer feedback to teachers and students to help them decide for potential corrective measures and/or adaptations and for improvement to bridge the gap between current achieved performance and determined objectives. Assessing (checking) and variability analyses is a key step in a continuous PDCA process as well as in further planning in establishing and assuring the quality of education at all levels.

Feedback is vital, it must help teachers and learners to close the gap between current and desired performance. It needs to be timely, specific, transparent, 'user friendly' and includes suggestions for ways to improve future performance. Good feedback is also tied to explicit criteria regarding expectations for student performance and making the learning process more transparent. The concept of feedback is a 'system' that operates with four components (Black, & William, 1998):

- data on student's actual level,
- data on reference level,
- a mechanism for comparing the levels,
- a mechanism used to alter the gap.

Teachers benefit from the feedback process. When providing feedback, teachers see examples of student work against expected standards, explore the meaning of the data and correctly interpret them, pay closer attention to what students do and do not understand well, understand student misconceptions, and are better able to adjust teaching strategies to meet identified student needs. School leaders should emphasize the importance of keeping the focus on teaching and learning, foster school-wide cultures of evaluation, engage in staff planning and action research, develop opportunities for teachers to provide peer feedback and support, ask teachers to refer to objective data on the impact of teaching methods on student performance, and in turn, have been asked to meet teachers' expectations for training and support.

It is important to focus students' attention on mastering tasks, rather than on competition with peers, and to help students feel safe to take risks and make mistakes in classroom. Children who feel safe to take risks are more likely to reveal what they do and don't understand, an essential feature of the formative process. Assessment is 'formative' when the information

gathered is actually used to alter the student's performance gap (*Formative assessment*, 2005). In the enlarged perspective of formative assessment the idea of remediation of learning difficulties (feedback + correction) is replaced by the broader concept of regulation of learning (feedback + adaptation) (Cardinet, 1977). His conception of regulation and adaptation was inspired by cybernetic systems analysis. A distinction was subsequently made between three modalities of regulation: interactive, retroactive and proactive regulation associated with formative assessment (Allal, 1979; Allal, 1988).

It was found useful to differentiate between two complementary levels of formative assessment. Level 1 concerns formative assessment that directly benefits the students who are assessed. Level 2 concerns situations where formative assessment data are used to inform teacher planning of future instructional activities proposed to new groups of students. Teachers are encouraged to carry out Level 2 regulations too, which in the long run can lead to systemic improvement of instruction. According to the OECD study, the use of data to inform teacher planning of future instructional activities for new groups of student (or at the policy level, for adjustment of policies) in the long run can lead to systematic improvement. That might be considered as a secondary level of formative assessment (Allal, Mottier, & Lopez, 2005).

Action implications

The development of the ALAT Tool will proceed in the following direction: it will be possible to analyse achieved objectives and performance standards with items analysis, to engage educators in reflection, discussion and planning focused on particular topic areas, taxonomy or levels of critical thinking issues, misconceptions and strategies, particular classroom practices and students behaviours, and to be less concerned with simplistic comparisons or ranking based on summative scores. The concept could be used also for primary schools where data are gathered from school- or teacher-based and national assessments.

It would be much better if both norm- and criterion-referenced uses of the results could be supported to complement norm-referenced interpretation of results with criterion-referenced interpretation. With criterion-referenced interpretation teachers can identify (diagnose) strengths and weaknesses in learning concepts to shape teaching for improvement. The procedures used by some examination authorities attempt this marrying of criterion- and norm-referenced assessment. Criteria are defined for some of the grade boundaries and marked student scripts at those grade boundaries in previous years are inspected in an effort to ensure that the definition and application of the criteria are consistent over years (McGaw, 2006).

The skills required for gathering and interpreting classroom, school or programme level data should be suitable for teachers and head teachers. From the methodological point of view, students' achievement could be presented in quantiles (quartiles) – with box plots. In schools, students' achievement is traditionally expressed in raw scores, or per cent, their knowledge is assessed with grades and grades are, as agreed, 'added up' to describe overall achievement. Comparing achievement expressed in average grades and standard deviation is sometimes not the most suitable way as distributions are not normal and the number of students in some schools, and mostly classes, is small. This is why it is better to represent the distributions to make comparisons of achievement, and we intend to produce an even more suitable method to compare different distributions of grades, point grades and overall achievement of populations, schools and classes, or groups of students. Comparisons of distributions have to be methodologically correct, illustrative and suitable for a widespread use among teachers.

The idea of school- or teacher-based evaluation is quite appealing because it involves school staff directly, incorporates local knowledge and, potentially, directly shapes school improvement. Newmann and Wehlage (1995) found that teachers and administrators in some schools formed a professional learning community, focused on student work and assessment, changed their instructional practice and support for learning in the classroom accordingly in order to get better results. Schools, in fact, do better when they pay close attention to external standards and corresponding achievement data (Fullan, & Watson, 2000). Focusing on students' work (assessment), changing teachers' instructional practice, and support for learning in the classrooms on a continuous basis all lead to effectiveness and improvement.

In Canada, all provinces and territories participate in a national programme to assess student achievement. Each province and territory receives its own results as well as an analysis by sub-test. Provinces may then conduct a secondary analysis to shape teaching practices. Due to the comprehensive and detailed nature of the data provided to schools, local school boards provide resources for each schools' leadership team to attend data-interpretation workshops. No data are given to a school whose leadership team has not attended the data-interpretation workshop. These workshops are clearly focused on the idea of assessment for school learning. Schools can use data to help set goals, allocate resources and plan interventions in areas that require improvement, as well as celebrate areas of strength and improvement (*Formative assessment*, 2005). Since 2001, the Department of Education in Newfoundland and Labrador has tested student performance in language, arts and mathematics on an annual basis. The Department advocates that the results of provincial tests be explicitly linked to school development. In some districts, schools are required to respond to the test data by completing a written analysis of how the school will use the data to improve the quality of instruction and which specific targets the schools sets for itself using the data analysis. Testing has changed the culture of communication about school achievement in Newfoundland and Labrador and has triggered communication at various levels.

The biggest challenge and an urgent action for the future is to promote the use of the Assessment of/for Learning Analytic Tool among teachers in schools and among other professionals. Teachers and schools' leadership teams should attend data-interpretation workshops for analysing the data, learn how to understand the meaning of the charts and how to judge the adequacy of the results, and how to improve their evidence-based work.

References

- Allal, L. (1979), "Stratégies d'évaluation formative : conceptions psychopédagogiques et modalités d'application" in L. Allal, J. Cardinet and P. Perrenoud (eds.), *L'évaluation formative dans un enseignement différencié* (pp. 153–183). Peter Lang, Bern,.
- Allal, L. (1988), "Vers un élargissement de la pédagogie de maîtrise : processus de régulation interactive, rétroactive et proactive" in M. Huberman (ed.), *Assurer la réussite des apprentissages scolaires ? Les propositions de la pédagogie de maîtrise* (pp. 86–126). Delachaux et Niestlé, Neuchâtel.
- Allan, L., & Mottier Lopez, L (2005). *Formative Assessment of Learning: A Review of Publications in French*. In: *Formative assessment – Improving learning in secondary classrooms*. Centre for Educational Research and Innovation. OECD.
- Black, P. & William, D. (1998). Assessment and Classroom learning. *Assessment in Education: Principles, Policy and Practice*, 5(1), 7-74.
- Broadfoot, P.M. (1996). *Education, Assessment and Society*. Open University Press.
- Burstein, L. (1980). The role of levels of analysis in the specification of educational effects. In R. Dreeben, & J. A. Thomas (Eds.). *The analysis of educational productivity. Vol. 1: Issues in microanalysis*. Cambridge, MA: Ballinger.

- Cardinet, J. (1977), *Objectifs éducatifs et évaluation individualisée* (second edition, Report NO. R77.05), Institut romand de recherches et de documentation pédagogiques, Neuchâtel.
- Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of students achievement or school effectiveness. *Educational and Psychological Measurement*, 57(3), 373-399.
- Creemers, B.P.M. (1994). *The effective classroom*. London: Cassell.
- Crighton, J. (2003). *Assessment: General Overview of International Practice*. Education Support Program, Open Society Institute, 1-22. Retrieved September 12, 2006 from http://www.soros.org/initiatives/esp/articles_publications/general_20030801/overview_20030801.pdf
- Deming, W. E. (1986). *Out of the Crisis*. Massachusetts Institute of Technology, Center for Advanced Engineering Study.
- Fertig, M. (2000). Old Wine in New Bottles? Researching Effective Schools in Developing Countries. *School Effectiveness and School Improvement*, 11(3), 385-403.
- Formative assessment – Improving learning in secondary classrooms* (2005). Centre for Educational Research and Innovation. OECD.
- Fullan, M., & Watson, N (2000). School-Based Management: Reconceptualizing to Improve Learning Outcomes. *School Effectiveness and School Improvement*, 11(4), 453–473.
- Gabršček, S. (2001). Access to Higher Education in Slovenia: assessment reform and re-creation of a national examination. *Assessment in Education: principles, policy & practice*, 8(3), 369–390.
- Gabršček, S., & Bethell, G. (1996). *Matura Examinations in Slovenia, Case Study of the Introduction of an External Examinations System for Schools*. Ljubljana: National Examinations Centre.
- Glossary*. (2006). National Foundation for Educational Research. Retrieved December 31, 2006, from <http://www.nfer-nelson.co.uk/glossary/glossary.asp?css=1>
- Guion, R. M. (1995). Commentary on values and standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 25-27.
- Hargreaves, A., & Fullan, M. (1998). *What's worth fighting for out there?* New York: Teachers College Press.
- Indicators on the quality of school education*. (2000). Activities of the European Union, Summaries of Legislation. Retrieved May 8, 2006 from <http://europa.eu/scadplus/leg/en/cha/c11063.htm>
- Kyriakides, L, Campbell, R J, & Gagatsis, A (2000). The Significance of the Classroom Effect in Primary Schools: An Application of Creemers' Comprehensive Model of Educational Effectiveness. *School Effectiveness and School Improvement*, 11(4), 501–529.
- Lee, V., Bryk, A., & Smith, J. (1993). Organization of effective secondary schools. In L., Darling-Hammond (Ed.), *Review of research I Education (Vol.19, pp. 171-226)*. Washington, DC:AERA.
- Linn, R.L., & Gronlund, N.E. (2000). *Measurement and Assessment in teaching (8th ed.)*. New York: Upper Saddle River, Merrill/Prentice Hall.
- Mc Gaw, B. (2006). *Assessment fit for purpose*. Paper presented at the 32nd Annual Conference IAEA – International Association for Educational Assessment. Singapore.
- Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School matters: The junior years*. London: Paul Chapman.
- Muijs, D., & Reynolds, D. (2000). School Effectiveness and Teacher Effectiveness in Mathematics: Some Preliminary Findings from the Evaluation of the Mathematics

- Enhancement Programme (Primary). *School Effectiveness and School Improvement*, 11(3), 273-303.
- Mullis, I.V.S., Martin, M.O., Gonzales, E.J., & Chrostowski, S.J. (2004). *TIMSS 2003 International Mathematics Report*. International Association for the Evaluation of Educational Achievement – IEA. TIMSS & PIRLS International Study Center.
- Newmann, F., & Wehlage, G. (1995). *Successful school restructuring*. Madison, WI: Center on Organization and Restructuring of Schools.
- Qualitative assessment of school education*. (2001). Activities of the European Union, Summaries of Legislation. Retrieved November 2, 2005 from <http://europa.eu/scadplus/leg/en/cha/c11038b.htm>
- Reynolds, D. (1998). World Class' School Improvement: An Analysis of the Implications of Recent International School Effectiveness and School Improvement Research for Improvement Practice. In A. Hargreaves, A. Lieberman, M. Fullan, & D. Hopkins (Eds.) *International Handbook of Educational Change*. Kluwer Academic Publishers. Dordrecht, Netherlands.
- Reezigt, G J, Guldemond, H, & Creemers, B. P. M. (1999). Empirical validity for a comprehensive model on educational effectiveness. *School Effectiveness and School Improvement*, 10(2), 193–216.
- Sheerens, J., & Bosker, R. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.
- Sanocki, T. (2001). *Student Friendly Statistics*. Prentice-Hall, Inc.
- Sliwka, A, Fushell, M, Gauthier, M., & Johnson, R (2005). Encouraging the Use of Summative Data for Formative Purposes. In: *Formative assessment – Improving learning in secondary classrooms*. Centre for Educational Research and Innovation – OECD.
- Topping, K.J., & Sanders, W.L. (2000). Teacher Effectiveness and Computer Assessment of Reading. Relating Value Added and Learning Information System Data. *School Effectiveness and School Improvement*, 11(3), 305-337.
- Zupanc, D. (2005). Standardi znanja v slovenski šoli [Knowledge standards in the Slovenian schools]. *Sodobna pedagogika [Contemporary Pedagogy]*, 56 (1), 124-147.
- Webster, B.J., & Fisher, D.L. (2000). Accounting for Variation in Science and Mathematics Achievement: A Multilevel Analysis of Australian Data – Third International Mathematics and Science Study (TIMSS). *School Effectiveness and School Improvement*, 11(3), 339-360.