DOI 10.2478/v10159-012-0003-y JoP 3 (1): 43 – 81

Lost in translation: Ever changing and competing purposes for national examinations in the Czech Republic[1]

D. Greger, E. Kifer

Abstract: In reaction to central control of schooling by the Soviet Union, the Czech Republic countered with what some say was the most decentralized system in Europe. While the political move to democracy was extraordinarily successful, there were numerous governments between 1989 and the present. The combination of the decentralized control of schooling and lack of continuity in the political realm in regard to education lengthened substantially the amount of time it has taken to mount national assessments. Those assessments, 5th and 9th grade and a high school leaving examination, are now on track but not without political and technical barriers.

Key words: Czech Republic, education policy development, testing, assessment characteristic, Maturita, misclassifications

Introduction

In this paper we describe features of the educational landscape of the Czech Republic and how they led to recent developments of national assessments. We focus on how the past and unique school circumstances in the Czech Republic are related to the development of those assessments.

There are three major parts of the paper: a chronology of educational changes since 1989; a way to view and categorize the proposed assessments; and, interpretations of the focus of the assessments as well as a description of their limitations as presently construed.

The assessments are a high school leaving examination and tests in the

[1] Acknowledgements: This paper is the output of research grant "Unequal school – unequal chances" (No. P407/11/1556) supported by the Czech Science Foundation (GA ČR).

5th and 9th grades. The purposes of the tests are not yet well-defined. And, a history of decentralized education and high turnover of educational personnel at the national level has led to long delays in implementing the assessments.

The paper is broad in the sense of trying to identify major factors that influenced the design and implementation of these assessments. It is narrow in the sense of being very specific in describing the assessments and their properties. Merging these two approaches provides a unique view of the process of creating assessments and implementing them in the Czech Republic.

Development of Czech education policy post-1989

The Czech Republic in 1989 moved from a totalitarian political system and centrally planned, state-owned economy to democratic governance respecting human rights, the restoration of private ownership and a market economy. These changes affected the education sector which, until then, was under exclusive central control. The political and economic transitions transformed the educational system. Those transformations can be divided schematically into four phases (Greger & Walterová, 2012).

Phase one – deconstruction

Deconstruction**,** the first and earliest phase of the educational transformation, lasted a few months after the political turnover in 1989. A short early period like this is typical of societies in transition, and is well-documented: Birzea (1996) labeled it *de-structuring* and Čerych, Kotásek, Kovařovic, Švecová (2000) termed it *a period of annulation or correction.* The main aim of this period was to redress the most visible shortcomings in education caused by totalitarian control.

Among the most important tasks of this first stage of transformation *were De-ideologisation* of the legal documents, including curricula programs, and *de-monopolisation* of state education. These facilitated setting up of private and denominational schools, stipulating that parents and students should be free to choose their schools and to pursue their educational goals. There was no emphasis on assessment and evaluation in this first phase of the transformation. There was, instead, in the centre of the debate, the liberation of content and methods of teaching and learning. Rigid political and ideological control in the system was replaced by a broad level of local school autonomy that Čerych et al. (2000) characterized as "unusually large and unparalleled in many western European countries."

School autonomy in this setting includes a wide range of policy options from curriculum determination to admission requirements and the content of examinations. Čerych et al. (ibid) argued that such school autonomy represented a complete departure from the old system and was the key factor in what became the bottom-up nature of the reform process in the first phase of the educational transformation in the Czech Republic.

Phase two: Partial stabilization

Kotásek, Greger, Procházková (2004) labeled the second phase (1991 – 2000) of educational transformation in the Czech Republic partial stabilization. After the most urgent and quickly made changes during the deconstruction phase, the partial stabilization period was characterized by gradual and incremental, legislative, organizational and pedagogical initiatives.

Despite a tendency toward keeping the "status quo," slow, deliberate/ partial adaptation to new conditions produced changes fostered above all by representatives of the school administration and conservative teachers. This period was still mainly one of bottom-up reform, where the main changes and innovations were promoted by individuals, local initiatives of teachers and Non-governmental organizations (NGOs). Reforms were mainly spontaneous, arising from the pedagogical issues and later based on operational, *"ad hoc"* measures. Partial stabilization is reflected in legislation by amendments to the communist era Education Act that was passed in 1984.

Among the key players in policy-making at that time were non-profit associations like the Independent Interdisciplinary Group for the Educational Reform (NEMES), an active teachers grouped (PAU – Friends for committed learning – see www.pau.cz) and a Group for Educational Alternatives – IDEA. These agencies and other expert teams prepared reform proposals without the state playing a leading role in policy development.

The Ministry of Education, Youth and Sports (MOEYS) in 1994 prepared a document, "Quality and Accountability", that was the first reform initiative from a governmental group with a national perspective. Although the report had no direct influence on education, it was the first attempt to formulate comprehensive policy with a long-term perspective. Thus the second half of the 1990's could be perceived as a turning point in policy formulation, with the state, represented by MOEYS, beginning to play a leading and major role in the process.

There were several other main sources of information that influenced the direction of change. Public opinion polls, analyzing the demand for schooling from different stakeholders, were conducted from 1995 till 1999 (see Kotásek, Greger, & Procházková, 2004; Walterová & Černý, 2006). Knowl-

edge of international and global trends in education was fostered by the active involvement of the Czech Republic in international large-scale studies of student achievement (e.g. Trends in Mathematics and Science Achievement -TIMSS 1995, 1998; International Study of Civics Education – CivEd 1999; Progress in International Reading Literacy – PIRLS 2001; Program for International Student Assessment – PISA 2000 and 2003). In addition, the Czech Republic participated in other Organization for Economic Cooperation and Development (OECD) projects, especially reviews of national policies for education in the Czech Republic (OECD 1996, 1999). The other driving force of internationalization was the negotiations and preparations for EU membership. This led to the preparation of an extensive strategic document, *Czech Education & Europe* (1999).

Phase three: Reconstruction

The second half of the 1990s was characterized not only by the partial adaptation and implementation of the changes required by the overall social transformation, it was also the preparatory period for the next (third) phase of transformation – reconstruction. The discussions about the future of national education were, according to Kotásek (2005), started in the second phase of transformation and then came to a head in the next reconstruction phase. Documents that influenced the changes included the White Paper (MoEYS 2001) and The Long-Term Plan for the Development of Education and the Education System in the Czech Republic, (MoEYS 2002). Those were followed in 2004 by the new Educational Acts (Educational Act – The collection of Laws on Pre-school, Basic, Secondary, Tertiary Professional and Other Education No. 561/2004, and Collection of Laws on Pedagogical Staff No. 563/2004).

Phase four: Implementation

According to Kotásek, et al. (2004) the last phase of transformation began in 2005 and continues today. It is a period of attempting to implement systemic reforms prepared in the previous reconstruction phase.

Analyses of the processes of transformation so far have been sketchy. There continue to be obstacles to the systematic understanding of the lively process of change, especially those that started spontaneously. Changes are still happening at the micro- or intermediate level, even though the macro level seems to be now in a final phase, perhaps ready for implementation. This implementation process, however, is not easy, especially for top-down reforms. Critics of the reforms (most often articulating their concerns in the domain of curriculum and evaluation) argue that the national reforms are

not well planned and, in particular, have not been explained and communicated to a wider public (parents and other stakeholders).

Politics and Educational Policy

A detailed explanation of the educational transformation has also not been sufficiently elaborated within the context of national politics. The preparation of systemic reform was made during the long period when the Social Democrats held power (even though it was a coalition government) lasting from 1998 till 2006. After the last election the leading party became the conservative Civic Democratic Party that has other priorities and the reform that was to be implemented by the Social Democrats is, itself, being reformed. Thus we might be observing the "*reforms of reform*", or what Birzea (1996) calls a *counter-reform*. The most visible 'counter-reform' is in the field of evaluation, where many measures prepared by the previous government and codified in law have either been postponed or are being gradually eliminated.

The process of educational transformation in the Czech Republic is similar in certain ways to what occurred in other countries and has been analyzed as the *tension between continuity and change*, a main feature of such transitions (Birzea, 1996). The current stage of development of education is either an implementation phase that requires a substantial amount of effort and time, or as a process of redefinition and reformulation of systemic reform. For both alternatives there are several obstacles to policy formulation or implementation, e.g. finances, management, but especially human resources (in case of evaluation, a lack of expertise in educational measurement, test designs etc.). The risk of reforming the reforms over and over again is thus the biggest obstacle to any change.

This brief outline of the transformation of education in the Czech Republic provides a bases and timeline to think about how and why the introduction of national assessments has taken more than 20 years. The story of the educational transformation in the Czech Republic portrays the struggles and shortcomings associated with introducing assessments without clearly setting non-contradictory, compatible goals.

The never-ending story: introducing national assessments in the 5th and 9th grade and standardized upper-secondary leaving examination

The Czech Republic is one of few OECD countries that until last year did not have national assessments or examinations. Even though the proposals

to reform the upper-secondary leaving examination (the *Maturita*) by introducing standardized state-administered tests began in the mid-1990's, and the first pilot tests were administered in 1998, they were formally required just last year. Results were reported, therefore, for the first time in the fall of 2011.

National assessments in the 5th and 9th grade (at the end of primary and lower secondary school) have a similar chronology. Their introduction followed long periods of discussion with some pilot tests in 2000's and an expectation of being fully introduced in 2013-2014.

The Czech System

Santiago, Gilmore, Nusche, and Sammons (2012) have an elaborate figure and discussion of the structure of the Czech educational system. To focus our discussion of the proposed new assessments, we present an abbreviated version of that structure.

The first level of education is pre-primary or nursery school for pupils age three to six. Almost all Czech children attend these schools where available, but they are not required to do so. Required schooling begins at age 6 and ends at age 14. This, the basic education has two stages. The first ends at grade 5 and age 10; the second ends at grade 9 and age 14. These two stages are the settings for two new assessments.

Although there are a number of options for secondary education, one with new assessments, the secondary schooling leaving examination (*Maturita*), is a subject of this paper. In general, passing the Maturita, usually taken after grade 13, enables a student to apply for admission to the university or equivalent education. Students in the Gymnasium, technical secondary schools and arts schools, take the examination. Students in a vocational track or school can get an apprenticeship certificate after 12 years of schooling, but there are also a few vocational programs leading to Maturita.

The proposed or newly instituted national assessments are focused on transition points in the school system. The 5th grade assessment is at the end of stage 1 or primary school. The 9th grade assessment forms a transition to secondary school. The *Maturity* examination is between secondary education and further or tertiary education. By being at those transition points, the assessments could conceivably be thought of as serving more than one purpose, something not considered desirable in testing circles. That is, they could be used to certify past performance or as a basis to select for additional education.

Prior assessments in the Czech Republic

Prior to the new assessment initiatives, such activities were conducted locally. Teachers and principals in secondary schools, for example, prepared their own high school leaving examination. Those examinations were written, administered and scored by local school personnel. They were usually composed of both written and oral sections but in some locations they were only oral. It is important to note these efforts because they are a piece of the history of the Maturity examination and apparently influenced greatly the design of the new assessment.

In addition to those mentioned above, outside companies produced examinations that schools could use if they so desired.

From local initiatives to national mandates and OECD influence

As discussed earlier, in the mid-1990s, after the revolutionary deconstruction phase of mainly bottom-up changes in education, the state became more involved in planning educational reform. At first, a number of strategic documents were formulated. Among the most influential ones were the OECD national education policy reviews undertaken in 1995 (see OECD, 1996) and in 1999 (see ÚIV, 1999). Many of the recommendations in the OECD reviews were taken into the strategic White Paper for educational development in the Czech Republic from 2001 (MoEYS, 2001).

While preparing their review of national education policy in 1995 OECD formulated 11 recommendations as priorities for action (OECD, 1996, p. 180). They grouped these recommendations into three broad fields:

- Improving the curriculum, structure and quality of basic and general secondary education.
- Strengthening the relevance, responsiveness and quality of administration of vocational and technical education.
- Implementing more effective means of administration, governance and management.

Five out of eleven recommendations were formulated for part A, among which two directly emphasized testing:

- Recommendation No. 1: *Developing instruments to assess pupil learning achievement in basic schools;* and,
- Recommendation No. 4. *Standardizing and differentiating the secondary-school leaving examination* (*Maturita*).

The two recommendations together and separately instrumentally changed Czech education. What might be the most famous and influential strategic paper on education in the Czech Republic, the so-called White paper (MoEYS, 2001) relied heavily on them.

OECD reviewers argued that in the Czech Republic "there is no appropriate instrument or set of instruments to assess and control the quality of education provided" and that "the traditional ways of controlling inputs ... can no longer substitute for the assessments of the outcomes of the teaching and learning process" (OECD, 1996, p. 181). Based on those assertions, they recommended that steps be taken

> to develop instruments to measure pupil achievement throughout basic school and to undertake a mandatory assessment of attainment at the end of ninth year. The assessment could take the form of a final examination in which an important place is given to an external measurement of attainment. (ibid)

Some possible approaches

The OECD reviewers emphasized that a change from measuring inputs to measuring outputs is a serious departure from previous activities and should be implemented with caution. They even suggested that external part of the examination should be proposed for use by schools but not imposed on them as a requirement. That is, schools could participate on a voluntary basis.

The OECD reviewers used a background report prepared by the Czech authorities (it forms the major and first part of the 1996 OECD book). The background report included still another critique of existing practices. Entrance examinations to upper-secondary schools that are organized by individual schools are fragmented and results from various schools are not comparable. What's more, the tests prepared by individual schools (or some schools use the tests of private providers), may not measure what was taught in basic schools, since their main aim is to differentiate among students for admissions (norm-referenced tests) and are, therefore, unfair. The OECD reviewers proposed to use examinations in the 9^{th} grade for measuring the quality of schooling, not for admission to upper-secondary level.

Even though OECD examiners did not provide details on testing (the whole recommendation is elaborated in just 1.5 pages), it is fair to assume that the main goal of the proposed examination was to measure the quality of education aligned to national standards (which they have said should be more fully elaborated). The ministry of education, however, added a goal to use the results of the 9th grade assessment for admissions to upper secondary school. It is, therefore, proposing to use the same tests for normative purposes. This is one of a number of examples of how the two opposing goals were actually set for the same assessment. This is highly controversial and considered to be technically unsound.

Another controversy linked to the assessment, and proposed directly by the OECD experts in 1996, refers to the use of assessment results for accountability purposes. Publishing league tables with results of schools (test-based accountability) is desirable, OECD reviewers argued, because "the publication of the assessment results, aggregated for each school but on a school-by-school basis, would help establish "quality maps" of the school system and encourage school directors, teachers and local School Offices, with appropriate support, to take action to upgrade those basic school demonstrating relatively low achievements" (ibid).

The proposal for test-based accountability is not surprising, since the New Public Management and accountability movement is generally dated to 1980s and the first research studies that challenged its effectiveness were of more recent (late 1990s). Since that time, however, there are several studies measuring the effects of test-based accountability, questioning its benefits (for a most recent review see Hout & Elliott (2011)) and stressing the negative unintended consequences it entails (like teaching to the tests and narrowing curriculum, inappropriate test preparations, affecting who is tested by exclusion, re-classification of students or retention of weak students in a grade). A more elaborate discussion of these side effects is contained in section 3.

The most recent OECD review on evaluation and assessment in the Czech Republic (Santiago et al. 2012) contains no mention or support for presenting league tables or test-based accountability. Rather what is proposed is the use of test results for giving feedback and for formative purposes. In fact, "assessments for learning" or formative assessment is stressed throughout the most recent OECD recommendations.

As we remarked in preceding paragraphs, the aims or purposes of assessments in 5th and 9th grade were not fully elaborated even though they included proposals for making schools accountable. The Czech authorities added another purpose, too – the use of test results for admissions to upper secondary level. We believe the two purposes, holding schools accountable and assessing student mastery of what is taught, is not achievable by one test. Again, according to testing experts, an inappropriate use of a test is to use it for two, perhaps divergent, purposes.

Clarity of Purpose

In the 2001 White paper, the purposes of assessment were not detailed; in fact, they were more general and in a broader context than earlier documents. The main argument was that autonomy for schools should be balanced by external evaluation and assessment, and more accountability

(MoEYS, 2001). The proposal for presenting league tables, however, was not mentioned in the White paper. The use of the results of assessments in 9th grade for admission to upper-secondary schools was kept and enthusiastically supported in both documents, the OECD one and the MoEYS one. Clear purposes for assessment and the potential conflicts among the purposes were never discussed. An assessment grid, like the one used to describe the assessments in next part of this paper was never completed and the requirements for tests not specified in a such a clear and concise way.

First Assessment Administrations

After the accession of the Czech Republic into the European Union in 2004, and with the use of funding from EU (European structural funds), pilot tests were administered in 2004 to 9th grade schools that volunteered to participate. The number of schools participating in the assessment increased over time from 2004 to 2008. The assessment in 5th grade started in 2005 and was administered also on a voluntary basis.

Each of the assessments included three tests: Mathematics, the Czech language and a student aptitude test supplemented by a student questionnaire. These projects were open to interested schools on a voluntary basis, and as early as 2008 there were about 1800 schools participating in the assessment in the 5th and 9th grades (about one third of all schools) accounting for 56,000 of 5th graders and 78,000 of grade-9 students. The financing for the project ended in 2008, however, and there was no immediate action undertaken to institutionalize these assessments on a regular basis as parts of a national evaluation system.

Governments change and so, too, assessments

There are several possible contextual reasons for the end of the financing. One is political, though not strictly ideological. The project was launched under the Social Democrats and when the financing ended, a new government was in power, formed as coalition of three parties with a Prime Minister from the conservative Civic Democratic Party. Neither the new conservative government nor the minister of education from the Green Party were specifically against the testing. Rather, it was a lack of strategic planning at the education ministry and lack of continuity among the governments that led to the collapse of the assessment initiative. Only during the last two coalitions led by the Social Democrats and their Ministers of Education was the ministry in power for an extended period of time (1998-2002 and from 2002-2006). This enabled them to prepare strategic documents and pass a new Education Act.

Since September 2006, and during the goverments formed by coalitions led by the conservative Civic Democratic Party, there were six ministers of education, with the average amount of time spent in the office being 11 months! What's more, usually the change of minister meant also the change of deputy ministers and even exchange of officers at the ministry and at lower levels of the hierarchy. This lack of continuity is, along with general opposition to introducing external assessments, a cause of the lack of funding for continuing assessments at 5th and 9th grade.

Even the OECD examiners suggested that change from measuring inputs to measuring outputs needed to be implemented with caution, argumentation and patience. To this concern can be added problems associated with goals of the assessments not being clearly stated. Goal statements, such as they were, changed over time, and often several contradictory purposes were mentioned.

In addition, there was nothing to suggest that the general attitudes of the Czech population were considered. Public opinion, as expressed in responses to a survey, and opinions of experts in education at the national level are split into two groups of equal size with different opinions. In the last public opinion poll on education in 2008, parents and the general public were divided in support for introducing an assessment of the whole population of students in the 9th grade, when 45% agreed, 42% was against and 13% of general public neither agreed nor disagreed with its introduction (Chvál, Greger, Walterová & Černý 2009). When asked further about how results should be used, of the group of respondents who favored grade 9th grade testing, more than 50% agreed with the use of results for providing feedback to schools, teachers, or parents. That is, the formative approach. Only about one third of those respondents, however, agreed that the test results should be included in students' grading or as one of the criteria for admission to high school, and even smaller group supported using the results for accountability purposes. The results were 28% agreed with publishing league tables and 27% agreed to use the testing results to evaluate local schools.

It should be noted that in the context of a post-Socialist country, the trust in the state or local authorities is generally low due to the strict control, with an ideological bias, over schools and peoples lives exercised under the rule of Communist party. This constellation of perceptions leads to lack of trust in local authorities being able to use, in an unbiased way, the assessment results to evaluate schools.

In phase one of transformation, the reforms in general were introduced as bottom-up initiatives. In that context, reforms strengthen the school's autonomy and eliminated the detailed prescribed curricula of the Soviet

era. In its place was a more general, and local, framework for curricula. For many experts and some parents the introduction of obligatory national assessments is understood as a step back into control and rigidity. Also, when the assessment for feedback purposes is offered to schools on a commercial basis by private providers (mainly the company Scio and Kalibro) the assessment results are not be used for purposes other than feedback to schools. The private providers are unlikely to give results to entities other than schools, thereby allowing comparisons among schools. So the state initiatives are more and more understood (and often also articulated by ministry officials) as the source of control, which still conveys negative connotations implicitly referring to the former non-democratic regime.

In addition to the split opinion on testing and cautionary tales of state control, there is a third argument opponents used against the implementation of the pilot tests. Both the private providers of the tests and the ministerial institute responsible for assessment , the Centre for the Evaluation of Educational Achievements (CERMAT), offered only norm-based assessments. This is in contrast to what was considered desirable – standards based assessments.

The OECD reviewers as well as White Paper on Education argued for development of curricular and evaluation standards; i.e. standards based assessments. But, that did not happen. The tests remained norm-referenced and their relationationship to what is taught in schools was unclear. Traditional reporting of results to general public was of very low quality, formal reports on psychometric qualities of the tests were not widely disseminated, and results of these assessments were not obviously used for either policy recommendations and formulation, or for monitoring the education system. Thus the state did little to prove the either the quality of the assessment or its usefulness. Because the recommendation for setting up standards was ignored, standards-based or criterion-referenced tests were not introduced.

These are the three reasons why the assessments have not yet been implemented. First is the fact that the development of the assessments is being rushed. The current conservative government, however, has set as goals for their term in office to introduce national assessment in 5^{th} and 9^{th} grade. The tests in mathematics, the Czech language and foreign language shall be IT-based (administered by computer), and composed fully of multiple-choice format questions. In reaction to previous criticism, the ministry started to develop evaluation standards upon which to base expected student learning or, perhaps, to develop the test. The time frame for development of these standards was however limited by the Minister of Education to only 6 months, with the result being standards of low quality.

A recent OECD (Santiago et al., 2012) review on evaluation and assessment in the Czech Republic came to the conclusion that

> the development of the standards is being rushed by the requirement for national tests to be piloted in 2011... Given the more immediate reason for their development, the standards may be more appropriately regarded as specifications for the national tests, rather than indicators of the quality of student achievement expected at different levels of education system. (Santiago et al., 2012, p. 10)

Also, the Review Team suggested the need for clearer articulation of the purpose of national tests and to recognize that they cover a limited range of competencies. The tests, as originaly announced by the Ministry, will likely be very "high-stakes tests". This will certainly arise if the test scores are used to evaluate schools and/or teachers. Overseas experience, especially in the United States, has demonstrated that there are serious negative side effects when national test scores of student achievement are used for these accountability purposes (*e.g.* "teaching to the test, "narrowing of the curriculum") (Santiago et al., 2012, p. 10).

The OECD reviewers nicely summarized the criticism made by Czech education experts (including the author of this paper) to the present situation in the overall development of the assessments. Instead of opening a discussion about clearly setting goals, developing proper standards, discussing how to limit the side effects and constructing high quality tests for useful feedback to schools and parents with pupils, the Minister of Education is arguing that there were years of discussion but no real action. So, he broke this chain and bravely introduced the assessments. Unfortunately, introducing the national assessment has become the goal in itself, rather than understanding the national assessment as a means to meet previously stated goals. The current minister and the Ministry did not formally define the purposes of the assessment, but in his public speeches he has already mentioned that the results of achievement tests will be used for evaluating school quality, for publishing results of the schools in league-tables, for admission of students to upper-secondary education, as well as for feedback of schools/teachers/parents/pupils.

After the widespread criticism of the high-stakes nature of the tests, the minister changed his rhetoric and stated that national assessment will serve only as a feedback to schools. Such rhetoric however, is one thing; what he will try to do may be another. The experience with already introduced upper secondary leaving examination taught that the reality is often different from

the rhetoric (even though the ministry promised not to present league tables of secondary schools, that is what actually has happened).

The New Maturita Test and Its Legacy

The 'New Maturita' assessment has a longer history than the 5th and 9th grades tests, and since it was finally administered in 2011, there is experience, albeit short, in the use of results and quality of the tests.

But first some history: the Old Maturita dates back to communist times and until recently was regulated, with amendments, by the Education Act of 1984. The Maturita examination and its content was the exclusive responsibility of the principal of each high school. The general guidelines provided in the Education Act specified, that in the most prestigious schools (gymnasium – general academic stream – attended only by approx. 20% of student population) the Maturita will be composed of two obligatory subjects (Czech language and literature, and a foreign language) plus two optional subjects (list of subjects from which the choice was made also by the principles of the schools). At the secondary technical and vocational schools the obligatory parts of the examination included only the Czech language and literature, then one optional subject provided by school, and finally a theoretical and practical assessment from one of the specialized subjects related to technical or vocational orientation of the study program.

The former, and typical, form of the Maturita was an oral examination in front of the board of examiners (school teachers). Usually one part of the Czech language examination was the essay test, written on one of the four proposed topics, and scored by the teachers of the school. Maturita in mathematics at some gymnasia could use subject-matter tests prepared and scored by mathematics teachers in the school. Hence it is clear that the content of the Maturita differed markedly from type of the school (gymnasium vs. technical or vocational school/ study program) and also between individual schools (e.g. different gymnasia).

Even though passing the Maturita is a *sine qua non* for applying to a university, the results of the Maturita, though one of the criteria for admission to higher education, have very little weight and practically no impact on admission decisions. Instead, universities develop their own admission tests (in some places complemented by interviews and other assessments) or use the tests of the private provider Scio (e.g., two of seventeen faculties at Charles University in Prague make use of the Scio tests; other faculties prepare their own admission tests). The Scio company offers subject matter tests in social sciences, mathematics, science and foreign language (English

and German) and scholastic aptitude tests. In the 2009 – 2010 academic year there were 40,000 applicants sitting for 100,000 Scio examinations.

Proposals for reforming the 'Maturita' date back to the mid-1990s and are also associated with the OECD review of Czech education policy. OECD reviewers (OECD, 1996, p. 132) pointed pointed out that for the 'Old Maturita' the failure rate was 1% (much lower than in other OECD countries), results were not comparable across schools, they were an inadequate measures of student performance, and that the results were of little use in admission to higher education.

In 1996 the report stated that

> gradually, the view is gaining ground in the central administration and in universities, that something must be done. Proposals for a new organization of the examination differ very widely. A national examination of whatever kind would probably be unacceptable in the present political context, in part because of the mistrust of interference from central authorities in the educational process. There is also fear from the educational community that a centrally set examination would negatively affect teachers' motivation to be innovative and that it would lead to 'teaching to the test. (OECD, 1996, p. 132)

In addition, however, it was asserted that universities were in favor of a more standardized Maturita, in order to use its results. It was stressed by reviewers that central actions in the 1990s and particularly national assessment proposals were very sensitive and delicate issues, particularly in post-Communist countries where central control was often too strict and widely misused. The reforms in 1990s emphasized decentralization, more autonomy to schools, freeing up the curriculum frameworks and giving teachers more freedom. In this context, OECD reviewers suggested that external part of the new Maturita should not be imposed on schools, but be offered on a voluntary basis.

OECD reviewers also recognized that the three different types of upper-secondary education (general – gymnasia, technical and vocational) are of varying quality with different student cohorts based mainly on the social background on the students (high SES students in gymnasium academic track and lowest SES students in vocational schools). The report said that the

> Maturita has to satisfy various needs, those of higher education and those of the labor market… In the opinion of the review team, only a Maturita comprising several variants, each with a different mix of disciplines, some more academically-oriented,

> some more focused on labor-market competencies, can meet this different requirements. (ibid)

In early 1990's this division made sense since only about one third of graduates from upper-secondary level schools transferred to higher education, and the majority of those were students from gymnasia. Now, however, with 55% of high school graduates applying to university, the situation has changed. Nevertheless, large differences in social composition of students within three different tracks remain the same and constitute a barrier to introducing one type of assessment for all. OECD examiners therefore stated that "differentiating the Maturita according to courses taken during secondary education would be just as important as progress towards comparability across schools" (OECD, 1996, p. 133). These insights turned into Recommendation No. 4, mentioned earlier, which emphasized the need for standardizing as well as differentiating the secondary school leaving examination.

Examiners proposed a new model to address three deficiencies of the old form: 1st) results of Maturita are not comparable across schools; 2nd) Maturita do not permit an assessment of the quality of secondary education; 3rd) Maturita is of little use for universities in their admissions. OECD reviewers formulated the following proposals for new Maturita

> ...reform should follow three key principles. Firstly, the new Maturita should consist of a combination of school-based assessment of achievement and externally-developed and universally applied common examination, the latter in order to allow and promote comparability internally and externally. Secondly, the new Maturita should involve the state, either at central or regional level, to a greater extent than is presently the case, so that the Maturita examination becomes an effective means of quality control for secondary education. Thirdly, the new Maturita should provide useful information about the quality and content of student achievement to university faculties and higher education institutions. It is recommended, therefore, that the Maturita be divided into two parts, one defined at the school level and another standardized at the regional or country level in each of the broad curriculum areas of secondary general and technical education. (1996, p. 185)

Just as in the case of 5th and 9th grade assessments, the OECD review provided an elaborate discussion of purposes and format of the new Maturita, compared to that of the White Paper (2001). The White paper specified only

that the New Maturita have a common, standardized, external part, organized by the state that provides comparable results among schools and students, and simplifies the transition to the universities. The White Paper (MoEYS, 2001, p. 97) added a new characteristic, which was not part of the OECD recommendation – the state administered common part of Maturita shall have two levels of difficulty. This 'innovation' created serious technical problems and complexities. Because of the large differences in types of students and curriculum coverage in the three different tracks (gymnasia, technical and vocational) the common part would be either too difficult for the majority of the students not attending gymnasia, or would cover just minimum standards that will not enable the universities to use the results of the New Maturita for selection purposes. Two levels of difficulty were supposed to be the solution to meet the two goals for new Maturita.

In 1997, pilot tests for the New Maturita were administered by the ministerial Institute for Information in Education, and then in 2006 when the Centre for Evaluation of Educational Achievement (CERMAT) was founded, the preparation moved there. Following the OECD recommendations schools participated voluntarily in pilot testing and every second secondary school took part in these trials.

Even though New Maturita was piloted in 1997, it was not codified until 2004 when the new Education Act (first Education Act in post-Communist period replacing the Education Act of 1984) was approved. The New Maturita was codified then and was supposed to come into legal force during the 2007-2008 academic year.

Protests of high school graduates, who questioned the fairness of assessment saying the evaluation standards were not known in time and they had no opportunity to learn the content, forced the parliament to postpone implementation from 2008 to 2010 (it was also political issue, since new elections were approaching and well-organized upper secondary graduates and their parents constituted a large part of electorate).

The two biggest competing parties – Social Democrats and conservative Civic Democratic Party- were both involved with the new Maturita. It became a political issue for at least two reasons: 1st) both main parties blamed the other that the Maturita examination was not prepared in time and could not be implemented as originally planned; 2nd) the costs of pilots were unprecedentedly high and grew geometrically. In 2009, another postponement was approved, with a new target date in 2011. The model for new Maturita was ever changing and several amendments to Education Act of 2004 were approved, specifying various details of this new examination. One of the most significant changes approved in 2008 consisted of introducing two levels of difficulty, as was proposed in White Paper.

After years of trials new Maturita was implemented in 2011. It contained two parts – a common, state-organized and administered part and profile part that was the responsibility of each school. We deal in this text only with the state-administered standardized part of assessment, so we will not describe the varying part of the new Maturita model the content of which is determined by principle of each school.

In 2011 two subjects were obligatory: 1st Czech language and literature, and 2nd either foreign language or math. Students could choose an examination of lower or higher difficulty. The Czech language exam is composed of a didactic test, an essay test and oral exam (prepared centrally, but administered and assessed according to specified criteria by school teachers). Students must pass both examinations. Students may choose optional examinations from subjects like biology, physics, civics, history, etc. In 2011 students could choose an obligatory lower level of difficulty in Czech language and, as one of the optional examinations, a higher level of difficulty in the same subject (Czech language). The examinations offered as optional subjects were only at the higher level of difficulty aimed at being used by universities for admissions purposes.

The full model, supposed to be introduced in 2012, shall consist of three obligatory exams: 1st Czech language, 2nd foreign language and 3rd one of the following three subjects: mathematics, civics/social studies, or information technology. However, this model was, by another amendment, postponed till 2013.

Experiences with first year of new Maturita

Assessment experts, as well as student groups, said the tests were of inferior quality. Also, students made pictures of the assessment sheets and published the tests on the internet.

CERMAT, instead of publishing sample items themselves, warned the students that they broke the authors' law and there might be legal consequences for those who have made the tests public. The lower difficulty Czech language test was highly criticized. A review of this test by university experts and teachers claimed that one quarter of the test items were flawed. CERMAT made the tests available after they appeared on the internet. But, they never published a report of the psychometric properties of the tests or the characteristics of individual items. Nor have they responded to the critical reviews. The lack of evidence about the quality of the assessments and failure to report the test characteristics produced scepticism about whether the published results were valid and reliable. The mere assertion of reliable and valid measures was not enough.

How the results were used received additional criticism. Critics argued that, though very expensive, the endeavor brought no new information at system level. Critics knew ahead of time the best results would be in the highest track, gymnasium, and the most problematic ones would be in vocational track, where in some (rather small) schools more than 70% of students did not pass. This result reflects mainly the nature of the student cohorts and produces no new information for system improvement.

A main fear of both schools and educational experts was that publishing school raw scores in league tables would be interpreted by general public as reflecting the quality of the school, rather than the characteristics of the student population. The Ministry of Education said it would not happen. Nevertheless the Ministry did publish league tables of the best schools – presenting 10 best schools within each of the three tracks and also by 14 regions, the three best schools. This produced further demands from the media and representatives of the regions for additional comparisons between schools. They requested the data and produced full league tables including every school in the region and its results.

CERMAT never explained how they produced the score for the school quality, when students were taking different exams, and two varying levels of difficulty, but also individual subject-matter tests with varying difficulty (the cut-off score for Czech Language was at 44% of correct answers and for mathematics at 33%). One wonders, what one score for each school actually means, and how it was computed?

The first year of experience with newly introduced external state-administered part of the Maturita did not help to convince either education experts, or the general public, that it was a valuable exercise. Low transparency of the assessment and its questionable quality raises doubts if the whole and lengthy process was not a pointless exercise. Even the OECD examiners (Santiago et al., 2012, p. 10) stated that multiple purposes of the school-leaving examinations raise some concerns" and they add following recommendation:

> National standardized tests (as well as school-leaving examinations) should be valid and reliable instruments, assess the breadth of learning objectives in the curriculum, and results should be used properly for their intended purposes... An independent working group with representatives from a range of sectors and organizations in education could be established to further debate the national test, monitor its implementation and conduct impact evaluations. (ibid, p. 17)

Certainly this is a way forward.

A more Precise View of the Proposed Assessments

As indicated earlier, the Czech Republic's school system is both highly differentiated and highly decentralized. It is highly differentiated in the sense that there are both different types of schools and special schools for different types of students. It is highly decentralized because, although a substantial amount of funding comes from the central government, principals and head teachers are responsible for a wide array of factors related to how the school operates and how it responds to initiatives from more central forces. Although there is a national framework that schools must follow, how they approach the framework is not proscribed. Hence, there is a good deal of autonomy for a school and substantial diversity between schools.

With such diversity comes a bevy of opportunities to assess students. Entrance examinations, for example, are used for students to move between each of the levels of schooling and, even under some circumstances, for entry to compulsory schooling. There are school leaving examinations and exit certificates issued by different types of secondary schools. In general, gymnasia and technical schools have school leaving examinations; vocational schools give certificates.

Until very recently assessments at all levels have relied heavily on oral examinations enhanced with additional written questions. The questions are formulated by the principal and teachers within each school. The variety of contents and lack of standardization is evident. In fact, universities stopped using the exit examinations because they were not comparable across schools and hence not comparable for students. Of course, there may be positive attributes of local testing that are being ignored.

Although schools use standardized tests for instructional purposes and to provide achievement information, movement between school types and exits from secondary schools are based on these, what could be termed, informal assessment methods. Uniform standardized testing in these contexts was and remains to large extent rare.

CERMAT

That is beginning to change. The Center for the Evaluation of Educational Achievement (CERMAT), instituted in 1999 and now a unit within the Ministry of Education, Youth and Sports, was designed primarily to produce a standard secondary leaving examination (nová maturita). In 2004, with a new education law, its charge was broadened to produce 5th and 9th grade tests in the Czech language, mathematics, an aptitude test, and study skills. By 2007 those tests were formally administered nationally with

schools participating on a voluntary basis. In 2007, almost 60,000 students in the cohort participated in the grade 5 assessment. CERMAT followed a highly structured approach to develop an assessment, including reporting results for students and schools. In addition it provides a portal to provide information for a variety of audiences and assistance to schools as they learned more and participated in the assessment.

For the 9th grade assessment that included a measure of general skills, or aptitude, as well as the Czech language and mathematics, it gathered information from about 70,000 students in about 1600 schools. The participation rate for schools was about 60%.

It is interesting that the implementation of the 5th and 9th grade assessments moved so much more quickly than that of the high school leaving examination, one main target of our paper. CERMAT was given the charge to develop the high school leaving assessment in 1999 and began tryouts as earlier as 2002. Yet it was not until the fall of 2011 that it was officially administered. However the new development in grade 5 and 9 assessments is that responsibility for its implementation was moved from CERMAT to the Czech School Inspectorate (CSI) that shall develop and administer the test. CSI, however, plans to subcontract other companies to prepare the tests.

Before getting to the saga of the leaving examination, it seems important to provide information about its structure, how it has been constructed, and the kind of information that is available about it. Secondary school leaving examinations may have varied designs and do not necessarily function in the same way (Ekstein & Noah, 1989; Robitaille, 1997). In order to make sense, therefore, of the Czech approach to the assessment, it is necessary to describe its characteristics and how CERMAT approached its task. For all the years of trying to legitimate the examination much other work peripheral to the actual examination was completed.

One way to describe the various features of a large-scale assessment is found in Table 1 from Kifer (2001). It shows the many dimensions that must be considered when constructing, conducting and reporting results of something like the school leaving examinations. By using this table, one can begin to draw inferences about decisions made by CERMAT as it responded to a number of these issues when constructing the assessment. It also serves the purpose of locating disputed areas of the assessment when we begin to discuss factors that apparently led to the decade long hiatus of the national implementation of the test. Further, it serves as background to a discussion of side effects or unintended consequences. As Scriven (1993) has noted in evaluation of programs „side effects often are the main event."

Table 1 The assessment grid

Assessment Grid				
Purposes/	☐ Achievement	☐ Accountability	☐ Instruction	
Functions:	☐ Monitor	☐ Certify	☐ Evaluate	☐ Compare
			☐ *Formative* ☐ *Summative*	
Measures:	☐ Content	☐ Other		
	☐ *One* ☐ *More than one*			
Targets:	☐ Student	☐ Class/Teacher	☐ School	☐ District/State/Nation
	☐ Elementary	☐ Middle	☐ Secondary	
Standards:	☐ Frameworks	☐ Content	☐ Proficiency	☐ OTL
	☐ Assessment			
Stakes:	☐ High	☐ Moderate	☐ Low	
	☐ *Rewards* ☐ *Sanctions*			
Outcomes:	☐ Status	☐ Growth/Change		
		☐ *Cohort* ☐ *Longitudinal*		
Assessments:	☐ Traditional	☐ Performance		
	☐ *Multiple Choice* ☐ *Norm Referenced*	☐ *Constructed Response* ☐ *Performance Events* ☐ *Writing on Demand* ☐ *Portfolio*		
Technology	☐ Calculators	Word Processors	☐ Adaptive ☐ Devices	☐ Other
Support:	☐ Students	☐ Teachers	☐ Staff	
	☐ *Tutoring* ☐ *Summer School* ☐ *Other*	☐ *Staff Development*		
Reporting:	☐ Students/Parents	☐ Class/Teacher	☐ School	☐ Public

Source: KIPER, 2001

In order to understand better the proposed high school leaving examination, we apply the grid to the structure of the proposed assessment. Later in the paper we will discuss how the structure of the assessment may or may not meet the needs of the goals of the assessment.

Purpose/Function

The first and most important dimension of the assessment is determining its purpose. Good tests can be powerful tools but it is unusual to find they cannot do more than one thing well. CERMAT decided to measure the achievements of students to certify their mastery of parts of the secondary curriculum. According to CERMAT's director they are not interested in making inferences about schools based on the aggregate achievements of the students. That is, they want their assessment to measure achievements not be a source for accountability. Great Britain, for instance, produces "league tables" by aggregating students' scores to compare schools. CERMAT's director does not want that and emphasizes the purpose of the leaving examination is ascertain what students know not to judge schools. That is, of course, a controversial issue that could conflict with the Minister of Education's views.

Measures

A large scale assessment could have one measurement or many. It could include affective measures as well as cognitive ones. It could assess many different areas but form a composite score. It could score examinations independently or create subtests and sum those.

The school leaving examination is clear about what will be measured and how. There are categories for the measures within the assessments. The first distinguishes between those tests that are compulsory and those that are optional. Presently the compulsory set includes Czech language and literature and either mathematics or a foreign language. In 2013 both mathematics and foreign language will be compulsory. The optional list includes an array of typical secondary school subjects; e.g., physics, biology, art.

The second distinction is based on the difficulty of the assessment. A student may choose either a basic assessment or a more advanced one. It is assumed that success on the more difficult test represents a higher degree of mastery of the content and skills taught in secondary schools.

Targets

Assessments might measure students solely to determine how well they perform in a certain area. But, they might also use the student scores to

make judgments about effectiveness of teachers or the relative status of schools. The choice of targets ties into the purpose of the assessment. Often aggregate student scores are used as accountability measures for teachers and schools.

As indicated earlier the targets of the assessments are students and there is a dispute about using results for any other purpose. Although the percent of students who are entering some form of tertiary education is increasing, the targets are pupils within the gymnasia and technical schools. That is, there has been no expansion of school types offering the leaving examination.

Standards

Standards are ubiquitous. There are standards that define content outcomes. There are frameworks for assessments and assessment standards. There are proficiency standards that are used to determine whether or not an examinee has "passed." There are Opportunity to Learn (OTL) standards to ensure that tests contain content that examinees have been exposed to.

The Czech Republic like other OECD countries and systems is intimately involved in the standards movement. Those content standards lead to test frameworks upon which the examination is based. Although it appears that the tests are carefully constructed, there is no formal statement of assessment standards serving as a basis for test construction. Also, although it is assumed that the tests cover content taught in the schools, there are no formal OTL standards. Although it would not be intended, the tests may favor some schools and their curricula over others. Without OTL standards for schools, the representativeness of the assessment is open to questions on a fairness dimension.

For the compulsory tests, there are established cut-points that determine a passing score. The implication is that there are, therefore, proficiency standards. We could find no information about how the cut-points were determined, that is, what method was used to establish them. That makes it difficult to determine whether they have consistent scoring and methods for determining passing scores or whether they are following agreed upon methods for creating proficiency standards.

Since the optional assessments are constructed locally, it is difficult to know what, if any, standards are being applied there.

Stakes

The stakes or consequences associated with testing are complex things. What is very consequential for one examinee may be of little perceived impor-

tance to another. The consequences for this assessment, however, are well-defined. The first is actually obtaining the stamp of approval. The second is being admitted to a university.

The stakes are high for students but, as indicated earlier, not for teachers or for schools. Passing the leaving examination is opening the door to higher education. There are opportunities to repeat an examination but such support is not emphasized (This is in contrast to some high school leaving examinations in the United States where students may take them as early as their second year in high school and can repeat them often in subsequent years).

There are also incentives to take the more difficult of the compulsory examinations and to take a greater number of optional tests. The more, the better, in terms of acceptance to university since these results of these assessments play a major role in admission to a number of universities in the Czech Republic.

Outcomes

One way to distinguish among outcomes in assessments is to ask what is being measured. For instance, if an assessment is given once, then one is measuring the status of the examinees – how much do they know and how does one describe the performance at that one time. If, however, one has a prior comparable measurement, it is possible to estimate how much an examinee has learned or grown. Growth or change measures emphasized years ago, were almost forgotten, but now are having a renaissance in large-scale assessments. Such measures are often associated with accountability policies.

This examination represents the culmination of learning during a student's

secondary education. It is a status examination (what does the student now know) not a growth or change measure (how much did the student learn in a given time period). There are no comparisons made from cohort to cohort and no growth or change measures available.

Assessments

There are a variety of types of questions used in large-scale assessments. They can be multiple-choice items, performance events, open-ended items (constructed response items), oral examinations or portfolios. The type of assessment should be related to what are considered to be the desired outcomes. Multiple-choice questions, for example, are very efficient ways to assess a broad spectrum of information. They are not so powerful in determining how an examinee reasons or gives justifications for responses. Essay

questions are very good ways to assess both knowledge and reasoning but present problems of agreement among raters about the quality of the essay. Such lack of agreement is an issue in oral examinations as well.

The type of assessment in the Czech high school leaving examination varies from test to test. The compulsory mathematics examination, for example, has about ½ open-ended or constructed response items and about ½ multiple choice items. The Czech language examination contains a multiple choice portion, a written examination and an oral examination.

The type of assessment for the optional parts is allowed to vary. Most schools, it appears, continue the tradition of having mainly oral examinations, but schools may and do have written portions, too.

Technology

Technology has huge effects on an assessment. Obviously using a calculator on a mathematics test can change the nature of the test. The same is true of using a computer and word processing software for an essay examination. Also, technology can be adaptive: students with various disabilities may be given accommodations on the test. An example would be a visually impaired student having a human reader or screen reader to read the examination for him or her.

Calculators are not allowed on the mathematics test. But there are other accommodations as part of the assessment. The accommodations range from giving more time to complete the examination to providing a visually impaired student with an aid, if necessary, to read the examination. Here, as in other parts of the assessment, CERMAT is operating within a consensus of good assessment.

There may be calls for more and more accommodations or as technology improves a portion of the assessment could be computerized. Such changes will produce questions of comparability across time and fairness of the examination.

Support

A new and different assessment carries with it a need to provide information about the assessment and support and training for those involved in the assessment. Substantial amounts of information about all aspects of the assessment are found in a dedicated and excellent website: www.novamaturita.cz . There one can find descriptions of the new assessment targeted on students and parents, universities and the media. Several sections are devoted to how teachers and examiners have been trained.

Careful scrutiny of the website produces a broad and deep understanding

of the assessment and the preparations taken to insure its usefulness. One thing that is missing in the support category is providing additional assistance to students who do not pass one or more parts of the assessment. It appears as though a student may retake one part of any examination, but there also appears to be no formal mechanism to provide additional support – tutoring, practice examinations – for that student.

Reporting

There a number of audiences for results of an assessment. In addition, to students, teachers and schools, there are the public and in this case divided governments. Each of the audiences wants the information for different purposes and each can influence how an assessment changes over time.

CERMAT provides information for each of these audiences. That includes score reports for students and schools, information for the media and examples of the assessment. Questions or items on the assessment are posted on the CERMAT website and information about scoring, and levels of proficiency are also evident.

It is necessary to have planned carefully what information is provided and how it is provided. Choosing released items carefully can moot such criticisms.

The 5th and 9th grade tests according to the grid

Rather than apply each dimension of the grid to each dimension of the 5th and 9th grade tests, we, here, shall highlight the major features of those assessments.

Purposes and Functions

Here as indicated earlier is a source of contention that has not been resolved. Is the assessment to be of what students know and can do and the results used to provide feedback to teachers and schools; or, is it to provide a basis for selection into the next level of schooling? If it is the latter, it is subject to one criticism of the use of test scores. One test score should never be used alone to make an important educational decision.

Targets

The issue of which targets is related to that described above. Both students and schools are targeted by these assessments. The issue is whether

one assessment can produce precise scores for students and adequate measures of school effectiveness. We think not.

Assessments

These tests apparently are going to be computer administered. That means that scores can be calculated quickly and results be available in close enough proximity to the assessment that they can be used for formative purposes. The downside is that they will be only multiple-choice tests. This, of course, limits what skills can be measured well.

Interpretation

In the first two sections of this paper, we described the nature of the assessment issues and the context in which they occur in the Czech Republic. In this section, we are going to comment on those issues. Our comments deal with the assessment in general, the 5th and 9th grade initiatives, and then the high school leaving examination.

The Assessment Initiatives in General

Clarity of purpose

Successful large, wide-spread state or national assessments depend on a number of crucial attributes. There must be clarity in terms of the purpose of the assessments. Are they to focus on instruction and achievement or accountability? Will they be used for future educational decisions (selection to higher levels of schooling) or will they certify a level of achievement? Will they be used to inform or to judge?

So far the Czech assessment project has been marked by a lack of clarity or agreement about the purposes of the assessment. This is particularly true of the 5th and 9th grade assessments where it is not clear how the results of the assessments will be used. Of particular importance is whether the results of the assessment will be used exclusively for formative purposes, feedback to the schools and teachers about what is being done well and not so well on this small number of content areas. Or, will these, arguably narrow, assessment results be used to track students into different types of schools. Assessments can fulfill the first purpose easily. For successful placement of students, additional information about academic performance and attitudes is necessary.

Classification accuracy

Each of the assessments, in one manner or another, classifies students and/or schools. Although there is available information about the desirability of having reliable measures, we could find no discussion of the accuracy of classifying students. For example, how will the decision be made about whether a student passes the high school leaving examination and how many mistakes of what kind will be made?

Reliability of test scores or acceptable classification accuracy is a result of the amount of error in the measurements. Note neither notion deals with the larger conceptual question of the desirability of the assessment. It is based on assuming that what is being done is defensible in broader ways. These are the right tests at the right time!

In terms of assessment programs and accountability systems that have proficiency standards (i.e. cut-points to make a go/no go decision about a test score), the right technical questions have to do with how good (accurate) are the classification decisions, not how reliable are the tests, even though the two ideas are obviously linked. Although a number of researchers have asked that kind of question, there is no consensus about what methods are appropriate to answer it. There are, however, strong approaches in a technical sense that are done to address such questions (Young & Yoon, 1998; Rogosa, 1999).

One study in the United States (Hoffman & Wise 2003) reported that about 2/3 of the school classification decisions, whether or not schools met their goals, were correct ones. In an accountability arena where schools can be rewarded or punished according to their scores, the question is whether that level of classification accuracy is sufficient. In the Czech context, the question is how large are the errors when forming a "league table." Classification accuracy is played out through the ranking systems. If the league tables lead to consequences (rewards or sanctions) other than public knowledge of a schools program, classification accuracy becomes even more crucial and becomes an issue that must be addressed in the construction of the assessments.

In the present Czech context, classifying student performance accurately is more important than accurately classifying schools. Each of the assessments leads to a yes/no decision and, therefore, issues of proper classification. In a second study, Huffman, Thacker and Wise (2003) looked at the accuracy of student classifications. Depending on the grade level and content area somewhere between 60 and 82 percent of students were properly classified. In this study the lower grades and reading were the areas where classification was highest.

It is important to note that a estimated classification rate of 82 percent means a large number of students are misclassified. If 50,000 students are tested and the classification rate is 82 percent, there will be 9000 students with incorrect labels. Figure 1, a dot plot, shows the number of misclassifications on cohorts of about 50,000 each. It is interesting to note that the size of the cohort in this study was about the same as the size of the cohort that took the Czech Republic's 5th grade assessment in 2008. So the results here are particularly appropriate.

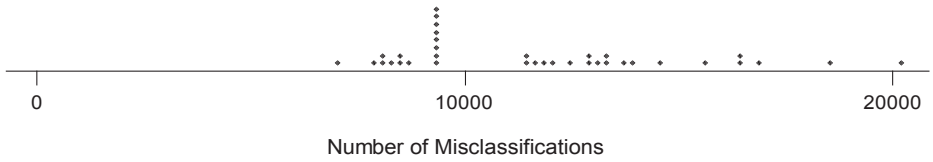


Figure 1: Student Misclassifications 2001-2002

Each dot on the plot represents an estimate of the number of misclassified students by subject area and grade level. For example, the dot furthest left on the scale is about 7000, the number of students who were misclassified by the 2001 11^{th} grade Social Science test. The whole dot plot contains, then, an estimate of the number of students for each of the subject area grade level combinations for the year 2001 and 2002 – the percent upon which the numbers are based are in Table 2.

We wish to make three general points about the above data display. First, these results show the limits of any educational test. These results are comparable to those of other studies of this type. The problem is clear: tests produce fallible measurements and fallible measures produce classification errors.

A mandatory high school leaving examination with a cut-point attached to it, and misclassifications of the magnitude found here, could be disastrous. As has been shown in the United States, there are huge controversies when erroneously scoring test items prohibits students from meeting graduation requirements which, in fact, they have met if the test were scored properly.

Third, there needs to be a way to communicate these results to various educational audiences to help them understand better what assessments can and cannot do. Persons should know that measurement error is an essential component of any assessment and that scores should be interpreted with that in mind.

The general research issue here is obvious, highly technical in some

aspects and extraordinarily important: What can be done to improve the precision of the measuring instruments used in the Czech Republic for areas that will be part of a high stakes accountability system? There are a number of questions related to that general one. Are school classification errors different depending on the kind (size, location, student composition, level) of school? Are student classification errors related to the background characteristics of students or the kinds of schools they attend? Each of these questions is related to the broader issues of equal opportunities for schooling, the possibilities for social mobility, and, in general, fairness.

Side Effects: The Main Event

For high stakes testing and accountability systems, just like other educational interventions, side-effects or unintended consequences are often the main issues. Among the negative side-effects are, for example, narrowing the curriculum, teaching to the test, consuming too much instructional time, intimidating teachers, and encouraging cheating. Some positive ones include focusing attention on standards, responding positively to criticism of schools and using test results as a means to get more resources for schools.

There remains research to be done to document or dismiss these criticisms and to understand better the broad implications of them regardless of whether the allegations are true or false. We should generate a richer research literature to document all effects, intended and unintended consequences, positive and negative, of high stakes tests and accountability systems. One piece that is missing in the Czech Republic plans is a systematic research agenda to understand better the consequences of its assessments.

In the United States, assessments are a lightning rod of the reform. It is arguable that those who wished to disrupt reform efforts found the high stakes assessment an easy target. How does one think about and explain these various effects?

More important, however, is how the assessments in Czech Republic may change the locus of control for education. The highly de-centralized Czech system is likely to change with the introduction of national tests. One would expect the central government through the Ministry of Education to exert more influence on the schools as the assessments become institutionalized. The testing tail is a powerful way to wag school dogs.

Finally it remains the case that the best information a parent can receive about the performance of their child is from the child's teacher. A score on a national assessment or nationally-normed test cannot begin to provide the depth and breadth of information that a visit to the child's teacher can give.

The 5th and 9th Grade Tests

It appears that these assessments are planned to do two things: 1) to produce scores for students to determine where they will be placed for subsequent education; and, 2) to judge the adequacy of the schools.

Those demands cannot be met with one assessment. Tests are narrow; schools are broad. A measure, for example, that estimates a student's mathematics proficiency samples but a very small part of a school's curriculum and even smaller part of what schools do. An assessment that is broad enough to sample well the curriculum may not be adequate to provide good scores for students.

Other large-scale assessments use a technique called item sampling when attempting to get a good measure at the school level. In that scheme, students receive different tests that, in the aggregate, represent broader samples of the content of a school's curriculum.

A second problem related to making determinations about schools is that a limited number of content areas are to be assessed. Schools teach more than mathematics and the Czech language. If one is to describe how well a school does in terms of what tests can measure, then all content areas should be sampled. Trying to do that, and to do it well, may be prohibitively expensive.

When high-stakes are associated with school outcomes, a number of negative side-effects accrue. As mentioned earlier, a set of common criticisms of high stakes assessments includes that their use leads to narrowing the curriculum both in terms of what subjects are emphasized and what in a subject area is taught (Ketter & Pool, 2001), teaching to the test (Shephard, 1989), and that too much time is spent preparing specifically for tests (Shephard & Dougherty, 1991; Herman & Golan, 1993). The net result is that higher scores may be an artifact: students can answer more test questions correctly but do not have any additional mastery of subject matter. Are there other measures that can be used to validate the score increases that are the inevitable result of high stakes testing?

There is probably more research in the United States on this issue than in other countries. In the United States scores from the Scholastic Achievement Test (SAT), the American College Testing Program (ACT), and National Assessment of Educational Progress (NAEP) often are used as criteria to legitimize gains on state assessments. Most such comparisons are either flawed, show that state score gains are not replicated in other test settings, or both (Linn & Baker, 1999, 2000, 2002; Koretz & Baron, 1998). In fact, Amrein & Berliner (2000a, 2000b, 2000c) argue that states with high stakes testing programs have on the average lower scores on SAT, ACT and NAEP.

There is a logic to findings that results from other tests do not confirm finding from findings from more local high-stakes tests. To do well in those situations, schools must emphasize the content and skills that are sampled by the state-wide assessments. In theory both the curriculum and the tests are built to reflect the same standards.

There is no such alignment with the SAT and ACT which were designed for a different purpose, to predict grades in college not to reflect a particular curriculum. In addition most states do not require students to take college entrance tests so it is not clear what kind of sample is represented by those who do take the test. NAEP purposely attempts to assess content that is not represented in any one particular curriculum. Hence, NAEP scores should be related to assessment outcomes only to the extent that the questions or items are sampling a common curricular domain.

The implications for the Czech assessments are clear: one important task is to define ways to establish the legitimacy of the testing. If the scores are just a reflection of a very narrowly tailored curricula experiences for students and do not provide wider interpretations, then any kind of ranking is suspect. The validity, usefulness, goodness, of the score interpretations are essential.

The High School Leaving Examination

The structure of the proposed policy for the high school leaving examination presents two major technical issues. The first is what is meant and how might one deal with assessments at two levels of difficulty. The second is how to score the leaving examination if it is composed of both a national piece and voluntary local assessments.

Psychometricians know how to create tests that are more or less difficult. They put more difficult individual questions on the more difficult tests. But difficult items may be only that. Take a simple example: To know whether a student can divide two numbers, one could ask the student to divide 20 by 5 or 10.7836 by 1022.432. Clearly the second question is more difficult. But, does successfully answering the latter question give more information about a student's mathematics achievement? Suppose the easy division were placed on the lowest level of difficulty test and the harder division question were place on the highest level. If a student taking the highest level test missed the question, but could answer the easy question, what does it say about that person's achievement.

Another way to make a question difficult is to write a question in a content area that some but not all students are exposed to. In mathematics it might be an easy question from a calculus course. If a student had taken

the course, the question might be easy. For a student who has not taken the course, the question is likely to look like a foreign language.

The general point is that it is not clear what inferences can be drawn from test score differences that depend on the construction of the test. Try-outs and transparency are a necessity! CERMAT must be much more open about how these tests are constructed and the results of the try-outs. In this case, of particular interest are the definition, construction and implementation of an assessment at two levels of difficulty.

The second issue is how to deal with scores that are defined by varied test components – the required and optional ones. In deciding how to aggregate a combination of scores, there are two general approaches. One is labeled compensatory; the other conjunctive. A compensatory approach creates a total score upon which to base the pass/fail decision. By using a total score, a high score on one test can overcome a low score on another test.

The compensatory model is most often used when each student is measured on the same components. That is not the case for the Czech assessment where students may have different components, and some of the components are local. So, it would appear that a conjunctive model will probably be used. A conjunctive model means that in order to pass the assessment, the student must pass each of the components. That may solve one problem of differing numbers of components but it creates another. Given patterns where a student may fail the required portion of the assessment but pass all of the optional ones or pass the required and fail an optional one, what should be the decision rule for passing or failing the assessment? These issues must be addressed in order to give the assessment credibility.

Some Final Thoughts

Amidst our discussion of the chronology, politics and structure of the new assessments runs another set of issues that need to be addressed. According to the experts, the Czech Republic had one of the highest correlations between the background characteristics (fathers and mothers educations and occupations) of students and their performance on international assessments such as TIMSS and PISA. This suggests that equality of opportunities, social mobility, and issues of fairness are intertwined with what is assessed and what should be done with assessment results.

This is particularly true if the early assessments, 5th and 9th grade, are used for the purpose of selecting those who will be placed in higher tracks or more select schools. Coming into play in these circumstances are the problems of using a single test score to make an important educational decision and the issue of substantial numbers of score misclassifications. In

addition, since those results will reflect only responses to multiple-choice prompts and will be based on a small number of content areas, the scores are not generalizable.

Students from higher social backgrounds will be favored in those settings with those measures. Rather than swooping in with a national test, a fairer approach would be to insure that each student is given ample opportunities to demonstrate their competencies. That would mean not only insuring equal access to the best learning opportunities but also the assessment of varied abilities with methods in addition to standardized tests.

There is irony when it comes to the high school leaving examination. There is little doubt that results of a national test would provide score comparability now not available with local assessment methods. The local methods might, however, produce fairer outcomes and more social mobility.

The argument is that local assessments, although less generalizable, are likely to produce better judgments of talent. If there are major differences between schools, it is possible for most students in one school to have higher national scores than most students in another school. Those scores will be influenced by a number of variables including the fact that the higher scoring students come from higher social class settings. But, the higher scoring students on a standardized test may not be the most talented students.

Because they are defined and conducted locally, the present local assessments are likely to identify the most talented students in the school, and do a better job than would be done with a one-shot national assessment. If that is the case, each school will have students who score at the highest levels and will be eligible for the best of further education. Given the environment, the student has excelled. There is no way the student can do better. That is, the local results may be both more representative but also be affected by a score ceiling.

This notion plays out in the United States at the University of Texas. Rather than admit students with lower test scores to increase diversity among undergraduates, a policy that could be declared unconstitutional, the university put into place a policy that admitted students unconditionally if they are in the top 10% of their high school graduating class. Because Texas schools are so differentiated by ethnicity, the new policy produces an entering class very similar in its background characteristics to that which would be admitted under an affirmative action policy.

Local decisions may be fairer decisions!

Whether the consequences of the new assessments in the Czech Republic

will produce more inequality is a question that should be investigated as the new assessments are phased into the system.

There are, of course, other major questions that can be addressed within the context of the

Czech educational system. Although OECD and other organizations push for a kind of testing Olympiad that seek to rank order what hypothetically is common to the varied participating systems, the Czech Republic, just as is true in other countries, has a history, a context, and a set of practices that are uniquely its. It is unlikely that any other OECD country began the 1990's with a more decentralized educational system. As assessments are introduced, will they be less successfully implemented because of the history? Or, will the changes, once implemented, be endorsed by all the major policy players.

The ramifications of a decentralized educational system, a supremely successful transition from Communist rule, competition among political parties for power and authority, provide a fascinating basis for a story of a changing educational system.

We watch with great anticipation!

References

Amrein, A.L. & Berliner, D.C. (2002a). High-stakes testing, uncertainty, and student learning *Education Policy Analysis Archives.* Retrieved March 24, 2011, from http://epaa.asu.edu/epaa/v10n18/

Amrein, A.L. & Berliner, D.C. (2002b). *The impact of high-stakes tests on student academic performance: An analysis of NAEP results in states with high-stakes tests and ACT, SAT, and AP Test results in states with high school graduation exams.* Tempe, AZ: Education Policy Studies Laboratory, Arizona State University. Retrieved March 24, 2011, from http://www.asu.edu/educ/epsl/EPRU/documents/EPSL-0211-126-EPRU.pdf

Amrein, A.L. & Berliner, D.C. (2002c). *An analysis of some unintended and negative consequences of high-stakes testing.* Retrieved March 24, 2011, from http://www.asu.edu/educ/epsl/EPRU/documents/EPSL-0211-125-EPRU.pdf

Bîrzea, C. (1996). Educational reform and power struggles in Romania. *European Journal of Education*, 31(1), 97–107.

Carnoy, M., Loeb, S. & Smith, T. L. (2000). *Do higher state test scores in Texas make for better high school outcomes*? Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Catterall, J. S. (1989). Standards and school dropouts: A national study of tests required for graduation. *American Journal of Education,* 98(1), 1 – 34.

Čerych, L., Kotásek, J., Kovařovic, J. & Švecová, J. (2000). The Education reform process in the Czech Republic. In *Strategies for Educational Reform: From Concept to Realisation.* Strasbourg: Council of Europe Publishing

Chvál, M., Greger, D., Walterová, E. & Černý, K. (2009) Testování žáků na konci základní školy a státní maturita – aktuální otázky současné vzdělávací politiky. *Orbis scholae*, 3(3), 79-102.

Clarke, M., Haney, W. & Madaus, G. (2000). High stakes testing and high school completion. *The National Board on Educational Testing and Public Policy,* 1(3). Retrieved June, 20, 2012, from http://www.bc.edu/research/nbetpp/publications/v1n3.html

Clements, S. & Kifer, E. (2001). *Talking back*. Frankfort, KY: Long-Term Policy Research Center.

Greger, D. & Walterová, E. (Eds.) (2012). *Towards educational change: The Transformation of educational systems in post – communist countries*. New York: Routledge.

Eckstein, M.A. & Noah, H.J. (1989). Forms and functions of secondary – school – leaving examinations. *Comparative Education Review*, 33(3), 295-316.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives,* 8(41). Retrieved March 24, 2011, from http://epaa.asu.edu/epaa/v8n41/part1.htm

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practices,* 18(4), 5-9.

Herman, J. L. & Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice,* 12(4), 20 – 25, 41 – 42.

Hoffman, R. G. (2002). *The Accuracy of students' novice, apprentice, proficient, and distinguished classifications for the 2001 and 2002 Kentucky Core Content Tests*. Frankfort KY: Kentucky State Department of Education. Final Report, HumRRO FR-02-46

Hoffman, R. G., Thacker, A. A. & Wise, L. L. (2000). *The Accuracy of students' novice, apprentice, proficient, and distinguished classifications for the 2000 Kentucky Core Content Test*. Frankfort KY: Kentucky State Department of Education. Final Report, HumRRO FR-03-06

Hoffman, R. G. & Wise, L.L. (2003). *The accuracy of school classifications for the 2002 accountability cycle of the Kentucky commonwealth accountability testing system.* Frankfort KY: Kentucky State Department of Education. Final Report, HumRRO FR-00-41

Hout, M. & Elliott, S.W. (2011). *Incentives and test – based accountability in education.* Washington, DC: National Academies Press.

Ketter, J. & Pool, J. (2001). Exploring the impact of a high-stakes direct writing assessment in two high school classrooms. *Research in the Teaching of English,* 35(3), 344 – 393.

Kifer, E. (1994). Development of the Kentucky instructional results information system (KIRIS). Guskey, T. (ed.) *High stakes performance assessment*. Thousand Oaks, CA: Corwin Press, Inc.

Kifer, E. (2001). *Large-scale assessment: Dimensions, dilemmas and policy*. Thousand Oaks, CA: Corwin Press, Inc.

Koretz, D. M. & Barron, S. I. (1998). *The validity of gains in scores on the Kentucky instructional results information system (KIRIS).* Santa Monica, CA: Rand Corporation

Kotásek, J. (2005). Vzdělávací politika a rozvoj školství v České republice po roce 1989 – 1. časť. *Technológia vzdelávania*, (3), 7–11.

Kotásek, J., Greger, D. & Procházková, I. (2004). *Demand for schooling in the Czech republic (Country Report for OECD).* Retrieved March 24, 2011, from http://www.oecd.org/dataoecd/38/37/33707802.pdf

OECD (1996). *Reviews of national policies for education. Czech Republic.* Paris: OECD.

Linn, R. L. (2003). Requirements for measuring adequate yearly progress. *CRESST Policy Brief – National Center for Research on Evaluation, Standards, and Student Testing*, winter 2003, 6, 1-4.

Linn, R. L. & Baker, E. (1999). Absolutes, wishful thinking, and norms. *The CRESST Line – Newsletter of the National Center for Research on Evaluation, Standards, and Student Testing*, Fall 1999, 1-8.

Linn, R. L. & Baker, E. (2000). Closing the gap. *The CRESST Line – Newsletter of the National Center for Research on Evaluation, Standards, and Student Testing*, Fall 2000, 1-8.

Linn, R. L., Baker, E. & Herman J. L. (2002). No child left behind. *The CRESST Line – Newsletter of the National Center for Research on Evaluation, Standards, and Student Testing*, Spring 2002, 1-6.

Linn, R. L.& Herman J. L. (1997). *A policymaker's guide to standards-led assessment*. Denver CO: Educational Commission of the States and the National Center for Research on Evaluation, Standards, and Student Testing.

Ministry of Education, Youth and Sports (MoEYS) (2001). *National Programme for the Development of Education in the Czech Republic. White Paper*. Prague: ÚIV, Tauris.

OECD. (1996). *Reviews of national policies for education. Czech Republic.* Paris: OECD.

Orfield, G. & Kornhaber, M. L. (Eds.) (2001). *Raising standards or raising barriers?: Inequality and high-stakes testing in public education*. Washington D.C.: The Century Foundation Press.

Robitaille, D. (Ed.). (1997). *National contexts for mathematics and science education*. Pacific Educational Press, Vancouver: Canada

Rogosa, D. (1999). *How accurate are the STAR national percentile rank scores for individual students? – An interpretive guide*. Retrieved June, 12, 2012, from http://www-stat.stanford.edu/~rag/ed351/drrguide.pdf

Santiago, P., Gilmore, A., Nusche, D. & Sammons, P. (2012). *OECD reviews of evaluation and assessment in education: Czech Republic. Main Conclusions.* Paris: OECD.

Scriven, M. (1993). Hard – Won lessons in program evaluation. *New Directions for Program Evaluation*, 58, 1-107.

Shepard, L.A. (1989, April). *Inflated test score gains: Is it old norms or teaching the test? effects of testing project.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. Retrieved March 24, 2011, from http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED334204&ERICExtSearch_SearchType_0=no&accno=ED334204

Shepard, L.A. & Dougherty, K. C. (1991). *Effects of high-stakes testing on instruction.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. Retrieved March 24, 2011, from http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED337468&ERICExtSearch_SearchType_0=no&accno=ED337468

ÚIV. (1999). *Priority pro českou vzdělávací politiku.* Praha: Tauris.

Voke, H. (2002). *What do we know about sanctions and rewards?* Retrieved June, 12, 2012, from http://www.ascd.org/publications/newsletters/policy-priorities/oct02/num31/toc.aspx

Young, J. M. & Yoon, B. (1998). *Estimating the consistency and accuracy of classifications in a standards-referenced assessment.* Retrieved June, 12, 2012, from http://www.cse.ucla.edu/products/reports/TECH475.pdf

Authors:

David Greger, Ph.D.
Charles University in Prague
Faculty of Education
Institute for Research and Development of Education
Myslíková 7
Praha 1
110 00
Czech Republic
Email: david.greger@pedf.cuni.cz

Edward Kifer, Professor Emeritus
University of Kentucky
College of Education
510 McCubbing Drive
Lexington KY 40503
USA
Email: skipk@uky.edu