7 The assessment and evaluation of educational institutions, school accountability

[Gábor Kertesi]

By the turn of the millennium almost every economically developed country understood the importance of measuring the performance of public education with comparable indicators, collected at reasonable costs. These indicators have to be linked to the activities of individuals making up the institutions and used to develop incentive systems that provide motivation for teachers, school principals and school providers to improve their performance. The efficient operation of public education institutions is contingent on having access to appropriate performance indicators and on linking this body of information to a well functioning evaluation and incentive system.

This paper is organised as follows. First, theoretical issues arising in connection with planning school accountability, assessment and evaluation systems are discussed. Next, the current Hungarian school evaluation system is described and the problems inherent in the system are identified. Finally, a plan is proposed for improving the system.

■ THEORETICAL FRAMEWORK

The problem: the importance of output based assessment. Public education is a complex system with an annual budget of about 500–600 billion Hungarian forints[1], most of which comes from central and local revenues and from private spending. A large number of social actors play a role in public education: 1.8 million students and their families, about 5000 educational institutions, 160 thousand teachers, and several thousand school providers (local governments, local government associations, foundations and churches). The public education "industry" can be described as a mix of several types of inputs and outputs. In the most general sense, the output of public education is the students' knowledge and skills in the broadest sense of the word, which they need in order to become successful members of society and to contribute to the development

[1] One Euro was equal to around 250 forints at the time of writing (May 2008).

of the country. The traditional view contends that the effectiveness of public education can be adequately assessed in terms of the resources used in public education: the number of teachers, the number of hours worked, the amount of grants per student, the buildings, classrooms, textbooks and computers used by education services, the number of teachers completing in-service training, the available curricula, etc. This view, however, relies on a false assumption, namely that "if a country spends a lot (or more than in the past) on public education, the system is guaranteed to function successfully (or more successfully than in the past)." Let us quote one of the main observations of the McKinsey report: "In fact, almost every country in the OECD substantially increased its spending on education over the same period, in addition to launching multiple initiatives to spend this money more effectively. Yet very few of the school systems in the OECD achieved significant improvements in performance. One study based on the results of national and international assessments showed that in many school systems performance had either flat-lined or deteriorated" (BARBER & MOURSHED, 2007, p. 10).

Educational inputs are not the right measure of educational effectiveness. It is not only the quantity of resources that matters but also their composition and the way they are used.

Educational inputs are not the right measure of educational effectiveness. It is not only the quantity of resources that matters but also their composition and the way they are used. Educational resources can also be wasted. The *efficiency approach* is different: we want to understand *the relationship between educational inputs and outputs.* The education system functions well if it functions effectively. We want to provide *feedback* for every stakeholder — parents and students, school providers, teachers and principals, as well as taxpayers — in order to help them in *identifying problems in the functioning of the educational system and improving performance.* What do we need to take into account if we want to design a well performing school assessment programme? There is a long list of problems that we need to tackle. First of all, appropriate indicators have to be found.

What kind of indicators shall we use? At first sight several inexpensive indicators are available: end of semester and year grades, exam results, grade retention, school continuation rates, etc. These data are, however, inadequate for our purposes as they do not allow inter-institutional comparisons. Better measures can be obtained from the labour market: returns to knowledge and skills acquired at a given school, i.e., employment rates, career advancement, wage and wage increase. This method, however, faces several practical obstacles: it would be rather costly to collect these kinds of data; there is no simple way of linking this information to the various levels of education (even less so to individual institutions). Even if this problem could be overcome, the results could only reach the educational institution involved after a considerable delay. Also this information cannot be directly used when plans for the improvement of educational practice are to be designed.

Another choice is the use of standardised tests which are designed to assess the basic components of individual competencies. This appears to be the most

promising method. The most appropriate tests are those suitable for assessing general skills that underlie overall learning abilities (i.e., the ability to acquire new knowledge of any kind). Examples include tests assessing reading literacy (the ability to understand texts, which is the most basic prerequisite for all types of learning), mathematical literacy and logical reasoning.

A standardised testing programme has *several advantages*: *a)* it allows inter-institutional comparison, *b)* the tests can be linked to universal benchmarks (e.g., at age *x* or in grade *y* students are expected to attain at least level *z*), *c)* the standardised test results can provide information which constitutes meaningful feedback for all stakeholders of the education system (schools, parents and school providers), i.e., information that helps them decide what is to be done if more than a pre-specified proportion of students fail to attain level *z* by the age of *x* or in grade *y* in a given institution. The information directly evaluates the institution, the proper locus of feedback and correction.

The use of standards based tests is also *not without problems*. There is enormous variation across individuals, which has a large impact on test results. The result of the assessment is therefore uninterpretable unless individual variation is controlled for. Individual level assessment is subject to a very large error term (the results are influenced by random factors). It is therefore desirable to aggregate the test results at school level. The aggregation of individual level results helps to reduce the measurement error but the volatility of aggregate data can have a significant distorting effect on the cross-sectional and longitudinal comparison of groups (especially for schools, school sites and classes with small student rolls[2]), where student composition may be highly instable at any one moment. Absences and other random effects may have significant consequences, and even relatively minor temporal changes (a student leaving or a new student enrolling) can lead to major temporal fluctuations in estimation results. These problems must be countered by a well-designed system.

A general theoretical framework that appears to be appropriate for the purpose is the human capital model, which takes into account the factors that have contributed to the attainment of students' skills.

How to measure the school's contribution to student achievement? Above all, we would like to highlight the importance of a theoretical framework. A general theoretical framework that appears to be appropriate for the purpose is the *human capital model*, which takes into account the factors that have contributed to the attainment of students' skills (measured by the test scores).

S_t denotes individual skill level (test score) attained in year *t* and the symbol *I* stands for the set of activities — involving the family, the broader environment or the educational institution — that may increase a student's skills either as a purposeful "investment" or as a "by-product" of some other activity. Let us further mark the time between birth and year *t* with the set of indices 0, 1, 2, ..., *t* (measured in number of years for simplicity). The problem then is described by the following simple model:

[2] Most poorly performing institutions belong to this category.

$$S_t = f_t(S_{t-1}, I_t),\ \partial f_t/\partial S_{t-1} > 0 \text{ and } f_t/\partial I_t > 0$$

The basic tenet of the model is that the level of skills attained in any given period is a positive function of the skill enhancing activities of that period and the skill level attained in the previous period. The model assumes that a higher initial skill level constitutes an advantage in acquiring new knowledge and that skill enhancing activities also have a positive effect on the performance measured in year *t*.

As a consequence of the recursive nature of the problem, the level of skills (the test scores) attained in period *t* is the result of the skill level given at birth and all the past and present skill enhancing activities (in the family, neighbourhood or school) that occurred between birth and period *t*:

$$S_t = f_t(S_{t-1}, I_t) = f_t[f_{t-1}(S_{t-2}, I_{t-1}), I_t] = F(S_0, I_1, I_2, \ldots I_{t-1}, I_t)$$

To achieve an estimation of the school's contribution to student competencies based on test results, we need to build a statistical model that is able to control for the effects of all past school and non-school inputs and all current non-school inputs significantly which may affect the student's performance measured in a given period.

Therefore, to achieve an estimation of the school's contribution to student competencies based on test results, we need to build a statistical model that is able to control for the effects of all *past school and non-school inputs*[3] and all *current non-school inputs* which may affect the student's performance measured in a given period. If these effects are not *controlled for*, the results will be biased since part of the effects will be ascribed erroneously to the school, while important school related effects may be ascribed erroneously to other factors.

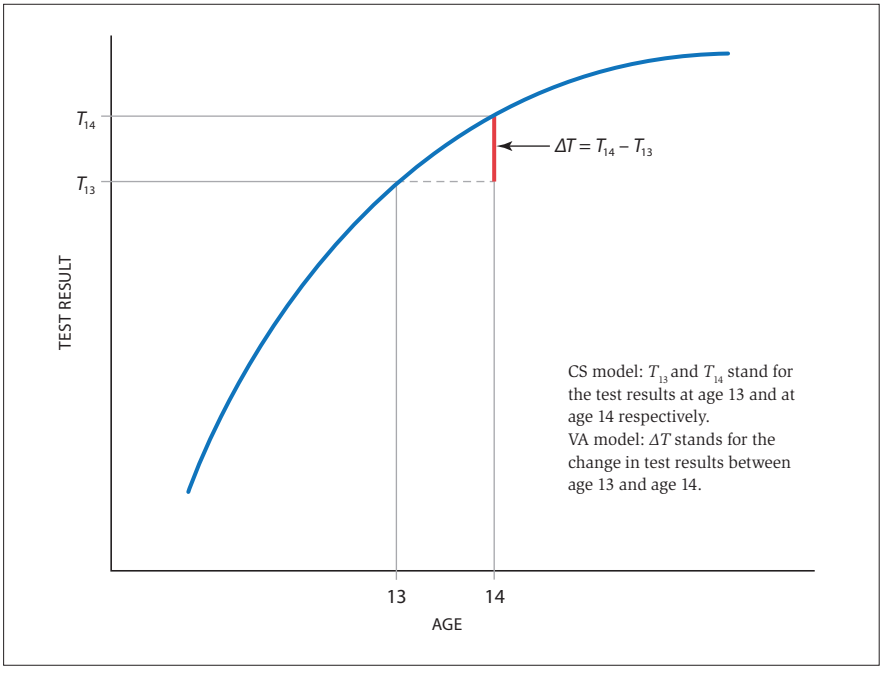
For simplicity, let us assume that the school's contribution is measured by some test results when the student is 14 years of age.

There are two alternative measurement strategies: we can use a cross-sectional (CS) model, which only makes use of information representing current effects, or we can use a value added (VA) model, which makes use of information from at least two consecutive cross-sectional measurements linked at an individual level. The CS and the VA methods rely on different outcome variables: the CS model — in our example — takes the test results attained at age 14 (T_{14}), while the VA model takes the *difference* between the results achieved at age 13 and 14 (ΔT). See *Figure 7.1*.

The cross-sectional (CS) model. This model takes the test results attained at age 14 as its dependent variable and only allows the effects of *current inputs* to be controlled for. To keep the discussion simple, disregard the difficulty that even current educational inputs cannot be measured directly but only by proxies such as parental educational attainment, employment status, the family's resources, cultural goods (the number of books) and similar data. *What kind of bias do we face when the school's contribution to the student's achievement is measured within this framework?*

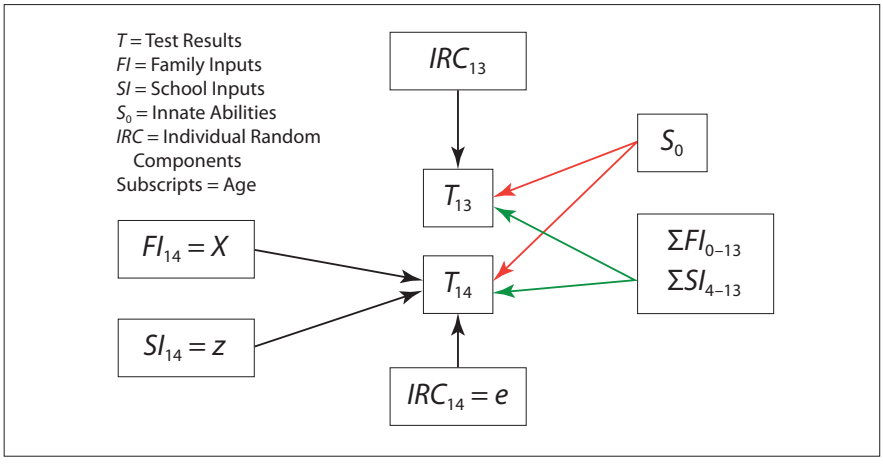
[3] As well as the effects of innate abilities.

[FIGURE 7.1]
Dependent variables of the cross-sectional (CS) and value added (VA) models



In *Figure 7.2* the inputs affecting the test results measuring the skills of a 14 year old are classified into four groups along two dimensions: current and past, family and school inputs. Current and past family inputs are denoted by FI_{14} and FI_{0-13}, current and past school inputs are denoted by SI_{14} and SI_{4-13}. Two further factors are included: the student's innate abilities (S_0) and individual random components affecting the test results (IRC_{14}).

[FIGURE 7.2]
The cross-sectional (CS) model



If the model were perfect, all of the factors listed above would be measurable. If this was the case, the school's contribution could be estimated by the following statistical model:

$$T_{ij} = X_{ij}b + \{\text{the effects of past family and school inputs and of innate abilities } (S_0)\} + \varepsilon_{ij,}$$

where i indicates a given student in a given school j, X stands for the proxies of the current family inputs, and ε_{ij} is the residual of this well specified estimate. Individual residuals can be decomposed into the school mean residual (z_j) and the individual deviations (e_{ij}) from that mean:

$$\varepsilon_{ij} = z_j + e_{ij}.$$

In the CS framework the school mean of individual residuals represents the school's contribution to the performance of students. Using the residual is the only viable solution to the problem of measuring the school's contribution. The more relevant the non current school input proxies included in the model are, the better can be our estimate. The problem is, however, that the cross-sectional model does not permit some important effects to be controlled for while empirical evidence[4] suggests that omitting these effects can significantly bias our estimate.[5]

The value added (VA) model. As an alternative, the value added model, offers a satisfactory — albeit not perfect — solution to the problem of omitted variable bias. To understand the logic of this measurement strategy let us return for a moment to the diagram of the CS model (*Figure 7.2*), where not only the test results at age 14, but the test results measured one year earlier, at age 13 are also displayed. Given certain conditions, it can be shown that if *panel data* of test results of consecutive years[6] are available, then the effects of the practically unmeasurable innate abilities[7] and of the difficult-to-measure past inputs can be eliminated (see *Figure 7.3*).

Two conditions must be met. It must be assumed that *(i)* innate abilities and past inputs are represented in linear form in the model and *(ii)* the parameters of *all* past effects from birth to the age of 13 (including innate abilities) in the

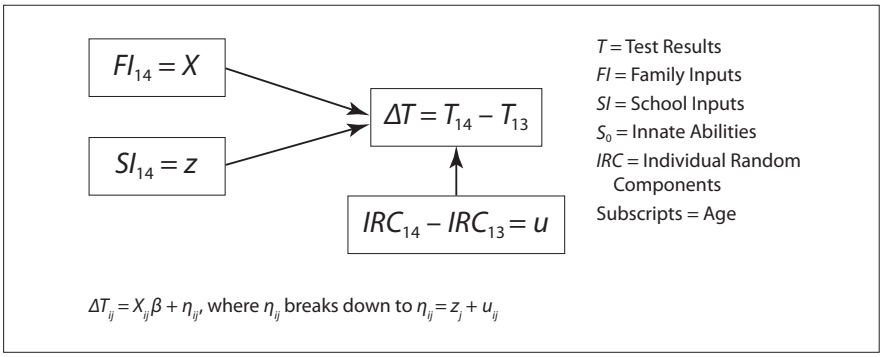
[4] See for instance HART & RISLEY (1995), LEE & BURKAM (2002), CUNHA, HECKMAN, LOCHNER & MASTEROV (2005).

[5] The bias may be mitigated (but not eliminated) by incorporating data on the student's family and schooling history in the CS model. This solution was used, for instance, by the Hungarian National Assessment of Basic Competencies programme in 2006.

[6] Every two consecutive years in Hungary as our assessment programmes cover students in grades 4, 6, 8 and 10.

[7] The only means of obtaining appropriate data would be by conducting very carefully planned long-term experimental panel surveys starting at birth.

[FIGURE 7.3]
The value added (VA) model



model for age 13 are equal to the parameters of similar effects in the model for age 14.[8] If these conditions are met, the effects of the effectively unmeasurable innate abilities and those of the difficult-to-measure past inputs can be eliminated by subtracting the equation explaining the test results at age 13 from the equation explaining the test results at age 14.

In the VA model, the school's contribution is once again estimated from the residual by taking the school mean of individual residuals (z_j) similarly to the method discussed in connection with the CS model, but the VA model allows a more reliable estimate since the effects of innate abilities and past inputs on the test results have been controlled for.[9]

Further theoretical and statistical challenges. There are three major groups of problems that must be faced: *a)* Since most of the theoretical variables (the "inputs") capturing the mechanisms of the theoretical model underlying the statistical model do not lend themselves to direct measurement, very careful consideration must be given to the choice of *measurable proxy* back-

[8] There may of course be objections to this latter assumption (see TODD & WOLPIN, 2003). The solution is not perfect but it is the best available.

[9] When the proposal was discussed by the Round Table, two arguments were raised against the VA model: *1.* that only those students can be included in this model who took the tests in both years (thus the size of the sample is likely to be reduced); and *2.* that the difference of two test results with a large random error is subject to an even greater random error (the errors add up). The answer to the first problem is that the assessment programme should be comprehensive in each year (as it is in grades 6, 8 and 10 as of May 2008) and the effects of mobility between institutions can be statistically adjusted with the help of the national student identification numbers. The problem of larger individual errors is a valid objection. It can be mitigated by using the school or site *averages* of individual "added values" and — as proposed here — by using *averages of consecutive years school averages* for the evaluation of a school's achievements. The question is, of course, whether it is the VA or the CS model that comes out as the winner when all the advantages and disadvantages of the two models have been taken into consideration. I believe that once these advantages and disadvantages have been weighed, the balance will unequivocally tip in favour of the VA model.

Since the school's contribution to student achievements can only be identified from the residual, no evidence can be obtained on the causes explaining why a given school's contribution to its students' individual achievements is found to be great or small. The mechanisms producing the effects must be explored in order to a deeper understanding of a school's contribution.

ground variables. *b)* Since the school's contribution to student achievements can only be identified from the residual, no evidence can be obtained on the causes explaining *why* a given school's contribution to its students' individual achievements is found to be great or small. The mechanisms producing the effects must therefore be explored in order to have a deeper understanding of a school's contribution. *c)* The residual based estimation of school contribution is highly sensitive to problems of sample size. The volatility of results caused by problems of sample size — especially for institutions having small student rolls[10] — concerns the *core* of accountability programmes. Let us discuss these problems one by one.

a) The choice of background variables is therefore crucial for the accuracy of the model whether it be a cross-sectional or a value added model. The relationships between the test scores and the background variables must be continuously analysed. Research and analysis are an indispensable part of the assessment and evaluation programme. No evaluation system of high standards can be delivered without this knowledge.

The identification of the mechanisms underlying good and poor school performance is one of the most important research tasks of the assessment, evaluation and accountability programme. It is essential to collect several relevant background information of schools, sites and classes in an effort to allow the heterogeneous causes underlying good and poor school performance to be analysed.

b) The analysis of school level residual effects[11] — the identification of the mechanisms underlying good and poor school performance — is one of the most important research tasks of the assessment, evaluation and accountability programme. It is essential, therefore, to collect several relevant background information of schools, sites and classes[12] in an effort to allow the heterogeneous causes underlying good and poor school performance to be analysed. It is far from immaterial, for instance, whether the good or poor school performance as measured by the residual stems from a shortage of certain resources or from an inadequacy in teaching practices or from the student composition. It matters because different causes call for different remedies.

c) The problems of sample size and the potential volatility of the results give rise to one of the most sensitive concerns with respect to the assessment of school contributions. The problem primarily concerns the assessment of the "performance" of institutions having small student rolls. *Figure 7.4* explicates the issue.

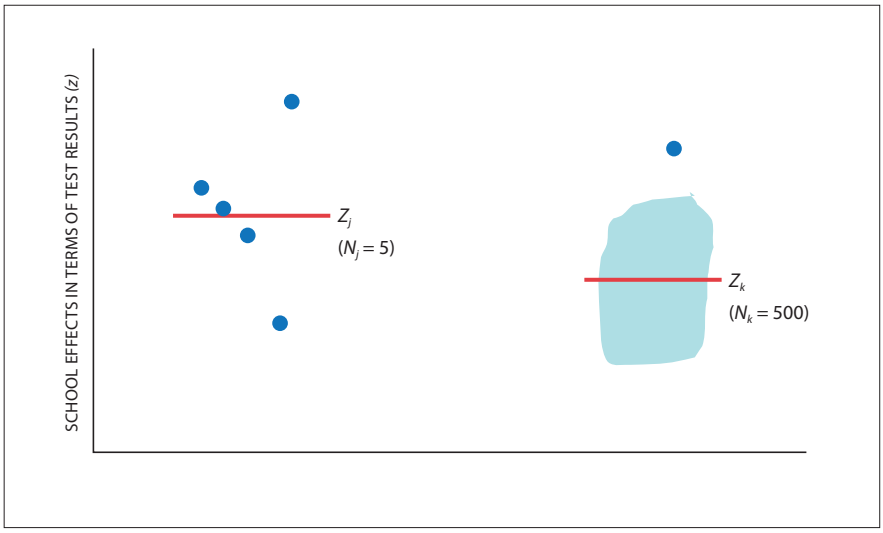
Two schools are compared. The figure displays the individual level residual effects — which are represented by individual dots and the gridded patch — and the school averages of the individual residuals (z_j and z_k). The latter val-

[10] Which incidentally appear to be the most prone to poor performance.

[11] We have no way of knowing what these are, we can only find out — and only within the framework of a well-specified model — what they are *not*. The better the model's specifications (the more known *non-school* factors we can control for), the more confident we can be that the school level average of individual residuals actually reflects the performance of the school.

[12] The collection of background details must be guided by theoretical relationships and thorough familiarity with available local and international assessment results.

[FIGURE 7.4]
The interpretation of school fixed effects and the stability of results



ues represent the "schools' performance." School *j* is very small (5 students) while School *k* is fairly large (500 students). What happens if a student who can display outstandingly good or bad performance does not take the test or misses one of the two consecutive tests? It is clear that the "average performance" of the small school is far more sensitive to even minor changes in the student composition than the larger institution. As a result, the "performance" measurement of small institutions is *highly unstable at any given point in time* and highly susceptible to *temporal fluctuations* in assessment results over longer periods. In other words small institutions are more likely to produce striking "improvement" or "deterioration," which may simply be a statistical artefact.

Sample size may give rise to a number of problems having a significant distorting effect on the assessment of an institution. *1.* The results of institutions with small student rolls are more sensitive to *random factors.*[13] *2.* The results of institutions with small student rolls may be more heavily biased should the *test results* be directly[14] or indirectly[15] *manipulated* in the school's favour. *3.* The smaller the institution, the more susceptible its assessment will be to *inter-institutional student mobility.* Given a small population of students, the temporal fluctuation of test scores may display a substantial "average improvement" due to the departure of a relatively low-performing student or the

[13] A barking dog distracts the students while they are working on the test. Student X is having a bad day or Student Y happens to have a lucky day, etc. — there is an unlimited number of possibilities.
[14] Students may receive help with the test from their teachers, for instance.
[15] Absences may be manipulated: students expected to produce poor results may be sent home. These students are more likely to be "off sick" on test days.

arrival of a relatively high-performing student and vice versa. Unfavourable changes to the student composition may similarly play a role in a slump in average school results. The smaller the student roll, the greater the fluctuation effected by changes of this type. *4.* Sample size may also pose problems in institutions having larger student rolls if the assessment of the institution's performance refers to *subgroup-specific standards*. If, for instance — as in the case of the school accountability programme in the United States — in addition to overall standards, the regulations specify standards applying to individual social, racial or ethnic groups in an effort to prevent institutions from meeting the required targets while neglecting the academic progress of disadvantaged minorities. These commendable considerations may, however, give rise to problems of sample size, which must be overcome just as in the case of small institutions.

We shall return to these problems later but let us now have a look at the uses of the school-level information obtained from the assessment and evaluation programme. What is the use of an indicator of "school performance"? This question leads us to the central problem of school accountability systems.

What sort of incentive system should be used? Let us assume that the assessment and evaluation system is well designed. Assessment methods have been thoroughly tested and have proved to be valid. Standards are clear and meaningful. We know what we expect of students in any given year of schooling. We also know that some of the institutions will probably fail to meet the standards. The question is, to what extent should the information provided by the assessment system be used to motivate the actors of the education system (schools, local governments) and to keep all stakeholders (families, students, taxpayers) informed? Two major classes of accountability systems have emerged worldwide. In one type the use of institution-level information gained from the assessment and evaluation programme is *strictly limited* to the provision of information for all involved and the general public *(soft accountability)*. In the other type of accountability system the results are associated with direct consequences or "high stakes": rewards are given and sanctions are imposed *(strict accountability)*.

Within the two classes, accountability systems display considerable variation in terms of the depth and breadth of information disclosure and the significance and type of rewards and sanctions. Among the sanctions imposed on persistently low-performing schools, one widely favoured intervention measure is the introduction of school choice for the students enrolled in the low-performing school in an education system where students are otherwise assigned to schools by districts.[16] (The provider of the low-performing school is required to bear the costs.)

[16] In the United States, for instance, the education system does not allow school choice by default.

Given that free school choice is a general feature of the Hungarian education system, our school accountability programme carries the mark of a strict accountability system by default even if no other rewards or sanctions are pledged.

This is highly relevant to our discussion because given that free school choice is a general feature of the Hungarian education system, our school accountability programme carries the mark of a strict accountability system by default even if no other rewards or sanctions are pledged.[17] Any positive evaluation of a school's contribution to student performance made public acts as an encouragement for families to opt for that school while any negative evaluation made public may deter families from sending their children to that school. These decisions bring about a direct gain or a direct loss of revenues for the school. *Whatever type of assessment and evaluation system is implemented in Hungary, it will necessarily have significant consequences because of the general availability of school choice.*

We do not yet have any experiences of the consequences of operating such a system. It is an important empirical research task to analyse schools' responses and refine the incentive system based on the results. The typical problems characterising school accountability systems are nevertheless well known from the international literature and the practical experiences of other countries, and we also have a reasonable idea of measures that can amend or at least alleviate these problems. The next section reviews these experiences and we shall also return to the problem of statistical validity.

What can be done about the problems typically characterising accountability systems? Four characteristic dilemmas will be discussed.

a) The complex nature of pedagogical objectives is at odds with the narrow focus inherent in assessment and evaluation systems relying on a few performance indicators (test results). The assessment system encourages schools to focus their activities on the chosen performance indicators while neglecting other educational objectives *(tunnel vision)*. A further negative consequence may be a practice of *teaching to the test* — a skewed teaching practice where students are mechanically trained to solve specific test problems at the expense of general skills development.

b) In genuinely problematic cases — such as small schools — the results are unreliable because of the statistical problem of *sample size*.

c) Schools may *manipulate test results* to their advantage.

d) As a result of the sensitivity of the method to sample size, institutions of different sizes have widely differing statistical odds of showing improvement or decline relative to a given performance baseline. That is, if educational institutions are evaluated in terms of a *standardised set of benchmarks*, *small schools* are more likely than large schools to be subject to rewards or sanctions, which is clearly inconsistent with equitability. What can be done about these problems?

[17] Other serious consequences are also planned to be included in the Hungarian system. See paragraph 99 of the Public Education Act and the Ministry of Education Act 3/2002 (II.15.) on the public education quality assurance and quality enhancement programme, which is currently under review.

If – in the hope of improving average test results – schools choose to train their students for the appropriate use of basic skills, teaching to the test no longer serves some futile purpose but directly contributes to the attainment of the required educational objective.

a) The problem of complex pedagogical objectives versus *performance indicator centred education (tunnel vision, teaching to the test)*. We propose the following countermeasures. *1.* The questions assessing students' achievements should focus on basic skills — such as meaningful reading — or higher order skills rather than on procedural skills relying on rote learning. If — in the hope of improving average test results — schools choose to train their students for the appropriate use of basic skills (for instance, to use reading skills for text interpretation), this type of teaching to the test no longer serves some futile purpose but directly contributes to the attainment of the required educational objective (the development of meaningful reading skills). *2.* The student and school assessment and evaluation programme should be gradually expanded to cover all major areas of competence. Assessments should be gradually rolled out to hitherto neglected competencies (scientific literacy, social skills, etc.). The present state of the system of education assessment must not be viewed as an invariable programme: major areas of competency should be given a balanced representation (the dependent variable should be seen as a vectorial one).

b) Small sample size can make the estimates on small schools unreliable. We propose the following countermeasures. *1.* The assessment programme must include all students. *2.* The evaluation should be based on individual panel data (the VA model). Any distortions caused by students' leaving or new students enrolling can thus be controlled for by statistical methods. *3.* School evaluations should be based on averages of average school results of consecutive years.

c) Test results may be manipulated by the school. Suggested countermeasures: A good solution to the problem is using individual panel data (the VA model) since *(i)* if panel data are available the bias due to purposeful absences can be measured by statistical methods and *(ii)* in an evaluation programme relying on value-added estimates (panel data) it is unproductive to boost school results by test manipulation since the results achieved in any given year are the base values for the following year and *artificially boosted base values reduce the odds of improved performance in the following year* (the ratchet effect[18]).

d) It is unfair to expect equal improvement of small and large institutions (or to impose equal penalties for a similar decline in their performance). Suggested countermeasure: standards should be adjusted to institution size. Large institutions should be rewarded for even relatively small-scale improvements.

[18] The ratchet effect: the phenomenon that performance expectations tend to be raised in an incentive system following the attainment of a markedly high achievement. This has the effect of penalising high achievement since further improvement in performance becomes more difficult to achieve and thus rewards become less accessible. See for instance, MILGROM & ROBERTS (1992).

The problem must be locally analysed and its causes must be revealed.

What should happen with low-performing schools? This is one of the core problems of school accountability systems. The first question to be settled is what we can expect from the disclosure of the assessment results of low-performing schools. The primary aim is to encourage *a local analysis of the problem in search of the causes*. Being able to localise poor performance is not at all equivalent to knowing how to handle it. Failure may be the result of several different causes. What we want to achieve by making the assessment results known to those involved is to encourage them to investigate the causes behind their failure and find the appropriate solution. An effective assessment and evaluation system is intended to offer an opportunity in this sense for the renewal of schools. To quote the aphorism by *Thomas C. Schelling,* "The problem with most incentive structures is *not* getting people to do the right thing. It's getting people to figure out what the right thing is to do." (Cited by ELMORE, 2004 p. 236)

"The problem with most incentive structures is *not* getting people to do the right thing. It's getting people to figure out what the right thing is to do."

Schools must adopt a culture of evaluating assessment results and they may need outside expert assistance with this task.

The core objective of a public education assessment and evaluation system is to transform schools into a problem solving organisation continuously reflecting on the outcomes of its own activities. To achieve this aim, schools must adopt a culture of evaluating assessment results and they may need outside expert assistance with this task.

There are several prerequisites to the task of identifying the causes of underperformance.

a) It is essential that the school have a teacher who is equipped to organise the work of analysing the results, that is, a teacher who possesses the knowledge and skills needed for the appropriate analysis and evaluation of the data and enjoys the authority needed to co-ordinate the activities of the teaching staff in this endeavour.

b) The entire teaching staff must be involved in the task of identifying the causes. This is important for two reasons. First, it clearly conveys the message to the local community that the school as a whole takes responsibility[19] for its students' results and second, it creates an opportunity to build a common approach, which in itself constitutes a first step towards a solution.

c) Schools may need external assistance with the task of identifying the causes. Central and local school governing institutions should undertake to ensure that schools have access to independent and competent professional help as needed.

d) The investigation into the causes may lead the school or the external expert advisor to conclude that the school's poor performance cannot be explained by deficiencies in the school's or its teaching staff's activities but appears to be the result of external factors, such as insufficient resources, the education poli-

[19] It would be unreasonable, for instance, to hold only the Hungarian language teacher responsible for the students' poor results in reading comprehension.

In Hungary we do not have an institution vested – with higher level powers in education matters than the local governments – which could offer an effective and binding institutional solution to the failures of the school system. The delivery of an efficient accountability programme is therefore contingent on the institutional restructuring of the national educational system.

cies of the local government or the spontaneous selection processes induced by the school choice system.

When we are faced with similar problems we must be aware of the fact that in present day Hungary we do *not have an institution – vested with higher level powers in education matters than the local governments* – which could offer an effective and binding institutional solution to such failures of the school system. The delivery of an efficient accountability programme is therefore contingent on the institutional restructuring of the national educational system with the aim of establishing such an institution. Chapter 10 of this Volume, where the institution structure and finance of education are discussed, presents a detailed proposal for the resolution of this institutional anomaly.

■ THE CURRENT HUNGARIAN SCHOOL EVALUATION SYSTEM

The current school evaluation system was introduced in Hungary in the 2001/2002 academic year. The main features of the system are the following.

Institution level evaluation relies on its students' individual performance, which is assessed with the help of standards based testing suitable for inter-institutional comparison. The tests currently cover two key competencies: reading literacy and mathematical literacy.

In addition to the total population of fourth grade students participating in a diagnostic assessment programme, all students in grades 6, 8 and 10 are mandatorily tested for these two competencies at the end of the academic year (on a given date everywhere in the country) as part of the normal school-year schedule. Only a small share of students with special educational needs are involved in the assessment programme.

At the time of testing, questionnaires related to family background are handed out to the students, which are optional to complete and do not include any questions that would allow the respondent to be identified. Usually about 80 per cent of students return completed family background questionnaires, which provide important background information for the evaluation of the test results. Public education institutions and their separate sites are also asked to complete background questionnaires on various school or site data, which provide similarly important information for the evaluation of test results. There are currently no unequivocal regulations with respect to the completion of these school/site questionnaires. Failure to complete them and inaccurate information provision do not have any consequences.

The completion of the tests is supervised by school-independent inspectors in only a very small proportion of schools.

Although the total population of students in the given years are required to sit the tests, the evaluation of the results and the *processing* of the available background information were *not comprehensive* up to, and including,

the test cycle in May 2007. Only a small portion of the tests were collected, coded, recorded and evaluated by the central agency. The central processing includes all the tests from a representative sample of 200 schools each for grades 4 and 6 and the tests of every second student in grade 10. The tests of all students in grade 8 are processed. The remaining tests are optionally evaluated by the schools.

The central processing of the tests is the responsibility of the assessment and evaluation department of the Educational Agency. Separate reports are prepared evaluating the results of all affected schools in the country at a school level, at the level of school sites, at a local government level and for the country as a whole.[20] These reports

a) present test result averages, national figures, the distribution of test results across school providers, schools and school sites, the distributions of test scores and skill levels;

b) contain calculations of school level test result averages with students' family backgrounds controlled for (this is the method used to estimate the school's contribution to student performance);

c) contain comparable data on the financial resources, facilities and social composition of schools and school sites.

Intertemporal comparisons can be obtained only by the comparisons of cross-sectional results of consecutive years. The reports are made available to every school and to every school provider. As of 2008, the Public Education Act requires school and school provider reports to be published on the web pages of the Educational Agency thus making them accessible to the general public.

The individual school and other (site and school provider) level reports on the competency tests form the basis of institutional accountability.

a) Based on the centrally processed test results of 6th, 8th and 10th grade students, the Ministry of Education and Culture regulations on quality assurance in public education annually specifies an upper limit to the acceptable proportion of students attaining the lowest skill level in a given year in a school.

b) If this upper limit is exceeded, the school involved must face serious consequences.[21] The first time this occurs, the school provider is obliged to call upon the school to draw up a plan of intervention within three months of receiving the call. The plan must detail the causes of the poor performance and set out a programme of enhancing the school's activities and improving student outcomes. If the school fails to reduce the proportion of low-performing students to a level below the specified limit as evidenced by the results of the third annual cycle of national assessment and evaluation after the call, the central Educational Agency — in fulfilment of its public education duties — calls upon

[20] See http://kompetenciameres.hu/2006.
[21] See paragraph 99 of the Public Education Act.

the school provider to prepare a plan of intervention within three months. The plan of intervention must be submitted to the Educational Agency for approval. The school provider is required to solicit the assistance of an educational advisory service or an education expert and the Educational Agency — in fulfilment of its public education duties — monitors the delivery of the proposed intervention programme.

■ DIAGNOSIS

1. In the current school evaluation system the evaluation of test results is limited to 200 schools for students in grade 6 and to half of the student population in grade 10, which makes the evaluation more susceptible to problems of sample size and thus unreliable. As a result of the partial coverage of central coding and recording, a substantial portion of all test results tend to be lost.

2. Only a small percentage of students with special educational needs (SEN) are tested as part of the assessment programme. This circumstance acts as an incentive to classify students as having SEN and thus exempt them for participation if they are expected to achieve poor test results.

The overall validity of the assessment results is not guaranteed.

3. Only a small percentage of classes participating in a given assessment cycle are supervised by a school-independent inspector at the time of sitting the tests. This level of supervision is insufficient to guarantee the overall validity of the assessment results.

4. Schools and school sites do not risk any sanctions by failing to complete background questionnaires. Since schools are publicly financed institutions, they should be under obligation to supply the required data.

5. The aggregate indicators characterising a school or site as a whole often mask problems which the assessment system is intended to reveal. Relatively large institutions may on the whole satisfy the requirements specified by the law while failing some subgroups of their students, such as *children of poor and uneducated parents*. This is a consequence of the absence of subgroup specific standards in the system.

6. The current system relies on school averages of residual test results with family background controlled for in measuring the institutions' contributions to student performance. This solution — which only controls for the impacts of *current* school and family inputs — is inappropriate since student outcomes are substantially affected by innate abilities as well as *past* family and school inputs, which cannot be taken into account in this model. If

The evaluation system could be improved by using students' identification numbers to link the results of consecutive biannual tests for each student, which permits the development of an evaluation model explaining changes in test results where the effects of past family and school inputs are controlled for.

these factors are not controlled for, the results are likely to be incorrect since part of the effects will be ascribed erroneously to the school, while important school related effects may be ascribed erroneously to other factors. An improved evaluation system could be developed by using students' identification numbers[22] to link the results of consecutive biannual tests for each student, which permits the development of an evaluation model explaining *changes* in test results where the effects of *past family and school inputs* can be controlled for. The school's contribution to student performance can then be estimated from the school average of individual residual effects similarly to the current model.

7. The standards used to evaluate school performance rely on information gathered within an *unjustifiably short period of time*: *the standards are required to be met by each of the separate measurements of consecutive years.* This evaluation system fails to accommodate the fact that the test results are exceptionally susceptible to fluctuations due to random factors, *especially for institutions with small student rolls* (see KANE & STAIGER, 2001, 2002). It is unreasonable to set the standards in terms of individual assessment results within a single year period. A more equitable reference point would be the *average* of consecutive yearly assessment results.

The implementation of the accountability system was accompanied by little effort to create the scientific resources needed for the comprehensive evaluation of assessment results and for refining the assessment programme. Without appropriate knowledge centres, however, schools cannot be reasonably hoped to master an evaluation culture empowering them to draw the appropriate conclusions from assessment results.

8. The implementation of the accountability system was accompanied by little effort to create the scientific resources needed for the comprehensive evaluation of assessment results and for refining the assessment programme. Without appropriate knowledge centres, however, schools cannot be reasonably hoped to master an evaluation culture empowering them to draw the appropriate conclusions from assessment results. The establishment and professional support of knowledge centres are indispensable for the maintenance of an extensive expert advice service needed for the pedagogical renewal of persistently low-performing schools.

9. The practical purpose of standards based testing and the information it provides has been subject to a great deal of confusion. The assessment and evaluation system regularly occasions misunderstandings in Parliamentary debates of the Public Education Act, in the Parliamentary Education Committee and among the education working groups of political parties. The diagnostic assessments designed for individual level educational intervention are regularly confused with summative assessments, which are designed to evaluate schools and not as a reference for individual level intervention. Law making processes in relation to the assessment and evaluation system are characterised by impatience and unjustified activism. Even though the Hungarian system

[22] As of 2007, the necessary legal conditions are granted.

is not yet sufficiently established[23] and its experiences have not been properly absorbed, policy makers bring forward a flow of *ad hoc* proposals on the fastest possible means of converting test results into school evaluations and on methods of penalising low-performing schools.[24] The tasks that should instead be given priority are the careful adjustment of professional standards, the refinement of assessment and evaluation plans, the scientific evaluation of assessment results, the establishment and support of professional working centres responsible for the development of programme contents, a training programme preparing a sufficiently large number of teachers for the task of test evaluation, the popularisation of an educational evaluation culture and the securing of proper financial resources for the assessment and evaluation programme.

10. The assessment and evaluation system is seriously underfunded. The usual budget sources cannot fully support a well designed assessment, evaluation and accountability system. The programme has been hampered by a constant shortage of resources from the very start of its existence (in 2001). The absence of a firm central budget commitment gave rise to an absurd situation in 2005, when as a consequence of the central austerity package, the just recently launched National Assessment of Basic Competences (NABC) supplying the data for the assessment and evaluation system had to be altogether cancelled for that year.

■ RECOMMENDATIONS

The tests should be centrally processed for each student.

1. The tests taken by all students in grades 6, 8 and 10 *should be centrally processed* (coded, recorded and evaluated) *for each student*.

2. To be able to follow individual student progress, the results of consecutive biannual tests should be linked for each student with the help of the student identification numbers, while at the same time respecting personal rights to privacy and data protection laws. The collection of test data should have a pan-

[23] This is the consequence of the complexity and novelty of the task and of a chronic shortage of funding. Some countries substantially more developed than Hungary took ten to twenty years to develop reasonably acceptable assessment and evaluation systems. Even these relatively well structured and appropriately funded systems are subject to continuous refinements and enhancements.

[24] This mistaken approach also surfaces in the current Public Education Act. Paragraph 99 of the Act — as was mentioned before — specifies serious short-term measures penalising low-performing schools even though the country completely lacks a network of experts who have the professional knowledge and capacity to assist schools in renewing their educational programmes, which is what low-performing schools would need.

el-like structure to ensure that the individual results of the assessments in grades 6 and 8 can be linked to the individual results of the tests taken by the same population of students two years later (in grades 8 and 10 respectively). In preparation for this programme the necessary student identification codes should already be stored in the next cycle of assessments due in 2008 to allow the data from 2008 to be linked to the results of the assessments in 2010 (and every two years thereafter) at an individual level. This step is currently under development by the Educational Agency.

3. Methods of assessing the competencies of students with special educational needs should be developed by 2010 in the framework of the Social Renewal Operational Programme of the New Hungary Development Plan (ÚMFT Támop). The assessment programme must then be permanently extended to the population of children with special educational needs.

4. The validity of assessment results should be ensured by the more extensive presence of inspectors. The supervision system should be extended step by step, i.e., the number of externally supervised test classes should be slightly increased every year. A reasonable target is having at least half of all test classes in grades 6, 8 and 10 supervised by an inspector in the 2012/2013 academic year.[25] In parallel with the expansion of the supervision system the class-level averages of the results of supervised tests and those of unsupervised tests should be subjected to comparative statistical analyses to reveal whether there are significant statistical differences between the two groups with all other conditions held constant.

5. The estimation of a school's contribution to its students' performance should rely on a value added model incorporating individual level panel-like time series data.

6. The New Hungary Development Plan funds should be used to implement a gradual expansion of the assessment programme to include further competencies, namely, scientific literacy, social skills and some other areas. Test contents should be continuously refined and adjusted. The assessment of reading literacy and mathematical literacy should be supplemented with tests measuring a different set of other competencies each year (as a pilot scheme).

7. Schools should be under a legal obligation to complete the background questionnaires about the schools and their sites.

[25] This is the level of supervision that seems to be practicable given that all testing is conducted on the same day. If the target was to have every testing session supervised by an inspector, the assessments would probably need to be spread over a few days, which would mean testing different grades on different days.

8. The available data should be analysed whether schools (and their different sites) meet the standards specified by the law for *the children of poor and uneducated parents as well*. This task crucially requires accurate records of the number of these students to be included in the Public Education Information System (KIR) database. Schools are currently required by law to keep these records. As part of the data collection activities preceding and preparing for each National Assessment of Basic Competencies, the Educational Agency should be given access to the data kept by the schools' on children of poor and uneducated parents enrolled in the relevant years. Within the period of a few years, the experiences of these measurements *may* lead to the introduction of a set of subgroup specific standards.

The schools' performance in relation to the benchmarks specified by the Public Education Act should be evaluated in terms of the averaged results of a number of consecutive assessment cycles.

9. The schools' performance in relation to the benchmarks specified by the Public Education Act should be evaluated in terms of the *averaged* results of a number of — say, three — consecutive assessment cycles[26] rather than in terms of the isolated assessment results of consecutive years. This method would substantially reduce the measurement error.

10. For the tasks of interpreting the final results of the assessment and for the planning of appropriate responses to them, knowledge centres specialising in assessment and evaluation, statistical analysis and social sciences should be established. It should be ensured that these centres have access to modern international theoretical and empirical evidence related to accountability systems. The knowledge centres should be granted support and reliable funding from sustainable budget sources. They could fulfil the function of expert advisory boards responsible for offering competent professional assistance to low-performing schools. The knowledge centres — preferably affiliated to major universities — could also function as training centres offering basic and advanced training for teachers to master the skills required for data evaluation.

Planned performance-based incentives should be piloted in schools within a relatively small continuous geographical unit and should be refined based on these experiences.

11. Prior to their wide-scale introduction, planned performance-based incentive programmes should be piloted within a relatively small continuous geographical unit (the schools in a single town or in a rural school association) with the voluntary and active participation of the selected local governments and schools. The plans should then be refined based on the experiences of the trials.

12. The central budget funding allocated for the assessment and evaluation programme should be substantially increased; the problem of permanent underfunding should be eliminated. The resources allocated for the system up to

[26] This method could be implemented by using three-year rolling averages.

and including the 2007 budget year are not sufficient for a modern assessment and evaluation system or for the implementation of the proposals outlined in the present study. It should further be ensured that the resources allocated for the programme remain stable and independent of the actual state of the central budget. A stable financial foundation should be secured preferably based on a per-student funding formula.

■ THE COSTS OF THE IMPROVED PROGRAMME

Up to and including the 2007 budget year, the Educational Agency responsible for administering the National Assessment of Basic Competencies programme and for evaluating the results had access to an annual budget of about 300 million Hungarian forints to implement the testing of the total student population attending grades 4, 6, 8 or 10 and to complete the evaluation of the results of 4th grade and 6th grade students from 200 schools each, the results of all 8th grade students and half of all 10th grade students (including the infrastructure and logistics of the entire process). The comprehensive processing of the complete set of data of grades 6, 8 and 10 (including coding and digitalising the data and evaluating the results) would cost about twice that amount: about *600 million forints* a year at current (May 2008) price levels.[27]

Assuming an intention to implement full-coverage central processing of the assessment results, external supervision covering 50 per cent of test classes in the above three years of study in 2012/2013 would cost about *250 million forints* at current price levels.[28] The costs of the stepwise expansion of inspector involvement depend on the pace of the expansion between 2008 and 2012. If, for instance, we set a target of 20 per cent supervision coverage for 2008, about 3000 inspectors will need to be delegated at a cost of about *95 million forints* at current price levels.[29]

With the figures presented in the previous two paragraphs added up, an annual budget of about *700 million forints* is required to sustain a modern assessment and evaluation system in the short term (in 2008),[30] and its maintenance would still not exceed an annual budget of *850–900 million forints* in the longer term (in 2012).

Further resources may be needed for the establishment and support of educational evaluation knowledge centres. These tasks can be partially funded from National Development Plan sources but their uninterrupted sustained

[27] Estimate approved by the Educational Agency.
[28] Estimate approved by the Educational Agency.
[29] Estimate approved by the Educational Agency.
[30] The Budget Act of 2008 allocates 700 million forints for the National Assessment of Basic Competencies due in May 2008.

operation also calls for reliable central allocations. Estimations of these costs are discussed in Chapter 9 of this Volume "The scientific foundations of learning and teaching."

■ LINKS TO OTHER PROGRAMMES

The recommendations outlined in this chapter share several points with the proposals discussed in Chapter 9 of this Volume. The modernisation of the public education assessment and evaluation system cannot fulfil its function unless the scientific and research base of pedagogical culture is strongly supported. A knowledge base is indispensable for the planning and successful delivery of professional intervention measures offering an — admittedly arduous — solution to the failures of the school system. In this respect our recommendations are also closely related to proposals aimed at palliating the school failures of children of poor and uneducated parents.

References

BARBER, M. ▪ MOURSHED, M. (2007). *How the World's Best Performing School Systems Come Out on Top*. McKinsey & Company.

CUNHA, F. ▪ HECKMAN, J. J. ▪ LOCHNER, L. ▪ MASTEROV, D. V. (2005). *Interpreting the evidence on life cycle skill formation*. NBER wp 11331. http://www.nber.org/papers/w11331

ELMORE, R. E. (2004). *School reform from the inside out. Policy, practice, performance*. Cambridge, MA: Harvard Education Press.

HART, B. ▪ RISLEY, T. R. (1995). Meaningful differences in the everyday life of young American children. Baltimore, London & Sidney: Paul H. Brookes Publishing Co.

KANE, T. J. ▪ STAIGER, D. O. (2001). Rigid rules will damage schools. *The New York Times,* August 13, p. A21.

KANE, T. J. ▪ STAIGER, D. O. (2002). Volatility in school test scores: implications for test based accountability systems. *Brookings Papers on Education Policy,* pp. 235–283.

LEE, V. E. ▪ BURKAM, D. T. (2002). *Inequality at the starting gate. Social background differences in achievement as children begin school*. Washington, DC: Economic Policy Institute.

MILGROM, P. ▪ ROBERTS, J. (1992). *Economics, Organization and Management*. Englewood Cliffs, NJ: Prentice Hall.

TODD, P. E. ▪ WOLPIN, K. I. (2003). On the specification and estimation of the production function of cognitive achievement. *Economic Journal, 113* (February), F3–F33.