A self-evaluation procedure for schools using multilevel modelling

Roel J. Bosker and Jaap Scherens

University of Twente, the Netherlands

ABSTRACT

For various reasons self-evaluation is becoming more and more important in schools. After highlighting the most common approaches to self-evaluation, it is argued that self-evaluation should primarily be based on the outcomes of educational practice. When designing a self-evaluation system one has to cope with two problems: how to define a fair school effect measure, and how to locate those practices that may lead to malfunctioning. It is shown how pupil monitoring systems can be used to construct a school monitoring system. The statistical aspects of such a system can be handled well by using multilevel statistical models. The proposed approach is illustrated using data on the development of pupils in mathematics achievement. The indicator proposed for self-evaluation purposes is compared with other indicators. Striking differences between indicators are found and discussed. Moreover, it is discussed how such a monitoring system could be modified to detect educational practices that lead to malfunctioning of pupils. Finally, the practical aspects of implementing the proposed system are depicted.

INTRODUCTION

One of the consequences of the Dutch national educational decentralization policy is the increased attention paid to quality control at the school level. With this increasing importance of the schools' autonomy, systems are designed and developed (and sometimes implemented) to help schools in their self-evaluation task. These include pupil-testing, management information systems, instruments for school diagnosis, etc.

In this article the attention will be focused on the application of multi-level models for the self-evaluation of the effectiveness of a school. The basis for this self-evaluation is the educational output, which is seen as the core criterion for educational quality. From this point of view there are two central problems in assessing a schools' quality:
1. how should we define a fair school effect measure and
2. how can we find those grades/teachers within a school that mostly affect the effectiveness of a particular school?

It will be demonstrated how pupil monitoring systems, that are being applied in about 35% of Dutch primary schools, can be used for the purpose of school self-evaluation. First, existing different approaches to school self-evaluation will be presented and discussed. Next the logic of using pupil monitoring systems will be described, and the statistical model to construct a school monitoring system will be introduced. This idea will be illustrated by using longitudinal data on mathematics achievement in lower general secondary schools (mavo). Eight different school effect measures will be defined, estimated and compared with each other. Furthermore it is discussed how to detect educational practices that may lead to malfunctioning. Moreover practical prerequisites of a self-evaluation procedure as proclaimed in this article will be mentioned.

* Adres: Faculteit Toegepaste Onderwijskunde, Universiteit Twente, Postbus 217, 7500 AE Enschede.
Dit is het laatste artikel behorende bij het themanummer.
A self-evaluation procedure for schools

APPROACHES TO SCHOOL SELF-EVALUATION

Evaluation is at the core of the concept of school effectiveness. The very question of judging the degree to which schools attain their primary objectives cannot be answered unless evaluative data, especially output data, are available. School improvement too – though this does not appear from the early improvement literature – builds heavily on the presumption that information on the current state of functioning of a school is available. In the more recent school improvement literature, to be associated with the OECD/CERI International School Improvement Project the recognition of the importance of evaluation data is clearly evident from the attention that is given to School Based Review (Hopkins, 1987). In more abstract terms, evaluation can be seen as the basic prerequisite for all types of control (management, planning, policy-making) (cf. De Leeuw, 1988).

There are three specific ways in which evaluation can be geared to the concept of school effectiveness:

1. (as defined above) evaluation of goal attainment to judge whether a school can be considered effective or not (usually this is done by comparing schools with each other);
2. the degree to which schools internally make proper use of evaluations (both at school and classroom level), which has been shown to be a favourable condition to school effectiveness; here evaluation, or “the evaluative potential” of the school is an explanatory variable with respect to the criterion mentioned above (under 1);
3. knowledge about the “process correlates” of educational achievement as established in school effectiveness research can be used in guiding the selection of relevant variables for educational evaluation systems, such as indicator systems (cf. Scheerens, 1990).

Currently, several approaches to school self-evaluation are being used. Each has a specific disciplinary background and a specific context in which the approach was originally employed, as is shown in Table 1.

Each of these approaches will be sketched briefly and strong and weak points will be discussed.

<table>
<thead>
<tr>
<th>approach</th>
<th>disciplinary background</th>
<th>context</th>
</tr>
</thead>
<tbody>
<tr>
<td>school-based review</td>
<td>social psychology, education</td>
<td>schools</td>
</tr>
<tr>
<td>management information systems</td>
<td>business administration, operations research</td>
<td>private industry</td>
</tr>
<tr>
<td>educational indicators</td>
<td>economics, educational statistics</td>
<td>macro-level applications</td>
</tr>
<tr>
<td>organizational diagnosis</td>
<td>management consultancy</td>
<td>private industry, public-sector organizations</td>
</tr>
<tr>
<td>pupil monitoring systems</td>
<td>educational measurement</td>
<td>(remedial) teaching</td>
</tr>
</tbody>
</table>
School-based review

School-based review depends heavily on opinions of school personnel on discrepancies between the actual and an ideal state of affairs in schools. In this way a broad perspective, in which all the main aspects of school functioning can be scrutinized, is possible. Usually, respondents are also asked to indicate whether a certain discrepancy should be actively resolved. In this way this approach to school self-evaluation seeks to gear improvement-oriented action to appraisal. The context of application usually is school improvement, which means that a school-based review is carried out when there is already a certain commitment to educational innovation.

Advantages of this approach are: a broad scope, a user-friendly technology, an explicit linkage between evaluation and action, and a high degree of participation (all school personnel takes part in the review). A definite weakness of school-based review is its dependence on subjective opinions and its (usual) negligence of “hard” factual data on school functioning, most notably output data.

Examples of procedures for school-based review are the GRID and GILS-systems (see Hopkins, 1987, and the SAS-system (Voogt, 1989).

School management information systems

School management information systems have been inspired by similar systems in private industry. Generally they consist of a careful modelling of information streams and information needs within a company, deciding which data should be available for purpose on a more or less permanent basis, followed by design and implementation of a computer configuration and software. Bluhm and Visscher (1990) describe a management information system as an information system based on one or several computers, consisting of a data-bank and one or several computer applications, which enable computer-based data storage, data analysis and data distribution.

A question that could be answered by means of such a school management information system would be: “to which degree has absenteeism decreased after the implementation of specific measures to fight absenteeism?”

Management information systems have a great potential for supplying important information on a routine basis. At present there are still a lot of practical barriers: one needs to have sufficient and adequate computer hardware and even when professionally developed software packages become available, quite a few specific maintenance functions must be carried out, while new routines and perhaps even functions to guarantee adequate data-entry should be developed.

Educational indicators

Although educational indicator systems are usually employed at the macro level (the level of national educational systems), for instance to describe the “state of education” of a country on a yearly basis, some authors have suggested applications at the school level (Teauber, 1987; Oakes, 1987; Scheerens, 1990). When applied at the school level, educational indicator systems typically will include “process” or “throughput” information, next to input, school-context and output data.

Results of school-effectiveness research studies are usually employed to select process indicators. The general idea of indicators is to provide an at-a-glance profile of certain important characteristics of an educational system.

This means that there is no aspiration to “dig deep”, while employing easily measured characteristics and so-called proxy measures. This feature is at the same time a definite limitation of the approach. Another “danger” is the use of process or throughput data as evaluation criteria, instead of explanatory conditions of educational outputs. This could easily lead to goal displacement, where the “means” in education are treated as “goals” in themselves. A technical limitation which might encourage this improper use of process indicators is the fact that the question of relating process and output indicators by means of formal statistical analysis has hardly been tackled for applied purposes. This problem will be addressed in other sections of this article.
Organizational diagnosis

As educational institutes (schools and universities) are made to function more autonomously, it is quite likely that they become more like private companies in their managerial and organizational characteristics. An example of this would be a stronger emphasis on strategic planning and on scanning the external environment of the school. It is therefore not surprising that approaches used in management consultancy are introduced in schools. Although these approaches, generally labelled as “organizational diagnosis” or “management audit”, usually depend on an external organizational consultant – they are also available for school self-diagnosis. In contrast to school-based review these approaches tend to be exclusively based on information provided by the management of the organization. So, when they are used without an external consultant they would appear to be somewhat like “management introspection”. A strong point of this approach is that it is likely to pay attention to issues that were kept largely unnoticed by the educational province, such as external contacts, anticipation of developments in the relevant environment, and flexibility in offering new types of services. The most important disadvantage remains, however, that this approach is not so easy to transform to a school-based application, without an external consultant.

Pupil monitoring systems

The focus of attention in this article is self-evaluation at the school level. Pupil monitoring systems operate at the micro level (class level) of educational systems. In the ensuing sections of this article it will be shown how this class of techniques can also be used for self-evaluation at the school level.

Basically pupil monitoring systems are sets of educational achievement tests that are used for purposes of formative didactic evaluation. An important function is to detect which pupils fall behind and where they experience difficulties.

Pupil monitoring systems have one asset which, in our opinion, is essential for all efforts to make school functioning more effective: the centrality of output data at the level of the individual pupils measured by means of achievement tests. If approaches to school self-evaluation neglect this type of data there is a risk that the information basis they supply for educational or administrative decision-making is faulty (see the earlier reference to the phenomenon of goal displacement).

SCHOOL MONITORING SYSTEMS

In the preceding section it was argued that the evaluation of the performance of a school should be primarily based on pupil progress in achievement. With primary schools being commissioned to refer less pupils to special education than before, schools try to assess pupil progress in order to detect, diagnose and remedy learning problems. For this purpose they use achievement tests. The Dutch National Testing Service provides tests, to be administered twice during the school year (Gillijns & Verhoeven, 1991). These tests meet the standards of Item Response Theory (IRT for short). This implies that they are very well applicable to the measurement of change (Moelands et al., 1989; cf. Rogosa & Willett, 1985). With the introduction of the common core curriculum in secondary education a similar system might be introduced based on the tests to assess progress of individual pupils with respect to the common standards. So from here on, the availability is assumed of a set of reliable and valid observations on the achievement of each pupil in school j on t different time points (t > 2): \( \theta_{jt} \) with \( \theta \) being a random variable with cumulative scaling properties. What we have here is the combination of a repeated measures design (within subjects design) with the nesting of pupils within a given school: the data are ordered hierarchically. In the statistical model to be presented, the basic modelling logic is that one wants to make full use of the data. As each sampling unit is sampled from a population, data from other sampling units (other time points, other pupils, other schools) will be used to improve estimates of pupil achievement, pupil progress, school output and school progress respec-
tively by applying Empirical Bayes techniques. The model described below is an application of a three-level model introduced by Raudenbush (1989, 1990) to study organizational effects on individual growth.

Since achievement is dependent on pupil characteristics like aptitude and sex the following prediction logic is used to model pupil progress (cf. Bryk & Raudenbush, 1987, p.148):

\[
\theta_{ij} = \pi_{0ij} + \pi_{1ij}t + \pi_{2ij}t^2 + \ldots + \pi_{(k-1)ij}t^{k-1} + \epsilon_{ij}
\]

in which \(\theta_{ij}\) is the ability-score of pupil \(i\) in school \(j\) at timepoint \(t\), \(\pi_{0ij}\) represents the base achievement level for this pupil at the start and \(t (= 0,1,2,3,\ldots,T)\) is time elapsed since the first measurement, with the remainder of the expression describing the growth curve characteristics, and the residual term \(\epsilon_{ij}\). According to the design of the CITO pupil monitoring system there are 14 measurement points in time (two tests taken at half a years’ intervals each grade from grade 2 up to grade 8). Expression (1) represents a polynomial of degree \(K-1\), that may be reduced to a smaller degree depending on the significance tests for \(\pi_{kij} = 0\) with \(k = 0,1,\ldots,K-1\).

In a graphical representation of the polynomial of the first degree we may find the growth curves as depicted in Figure 1:

In this example the two pupils differ in their initial achievement as well as in their growth rate. Explaining these two differences can be achieved in the second equation:

\[
\pi_{kij} = \pi_{k0j} + \beta_{kij}x_{ij} + \epsilon_{kij}
\]

which says that the variation in the growth curve characteristics \(\pi_{kij}\), with \(k = 1,2,\ldots,K-1\), of a pupil \(i\) in school \(j\) can be accounted for by a pupil variable like 'aptitude': smart pupils, for
A self-evaluation procedure for schools

instance, may have a high initial achievement status and a steep growth curve. It is important to stress this point, because this specification is necessary to distinguish school from pupil effects. What is special about this formulation is the allowance of random parameters in this between pupil model: \( \pi_{k0j} \) are the common elements of the growth curve characteristics of all pupils in school \( j \); furthermore the regression of pupil growth curve characteristics on pupil background variables may vary between schools. Aptitude, for instance, may affect the growth of pupils in some schools (e.g. schools using ATI-principles) to a much lesser degree than in other schools. In order to keep the discussion as simple as possible, it is assumed that the random variable \( \pi_{k0j} \) is of prime interest. Furthermore it is assumed (for simplicity sake) that the variance of the parameters \( \beta_{kj} \) is 0 (\( \beta_{kj} = \beta_{k} \) for \( k = 0,1,...,K-1 \)). The \( \pi_{k0j} \) are school specific, and they can be expressed as:

\[
(3) \quad \pi_{k0j} = \pi_{k0} + \gamma_k z_j + u_{kj}
\]

In words: the mean growth curve characteristics of school \( j \) consist of school specific deviations \( u_{kj} \) from the growth curve characteristics across all schools \( \pi_{k0} \). The mean growth curve characteristics of a school can be explained by a school level variable \( z_j \) like ‘denomination’ for instance.

How can this hierarchical model be used for self-evaluation?

First of all, things may be simplified by assuming that there are no background variables. Then one is left with the following simple expressions:

\[
(4) \quad \pi_{k0j} = \pi_{k0} + u_{kj}
\]

which indicates how to draw a line through all the observations of all schools, and how individual schools deviate from this line.

\[
(5) \quad \pi_{kij} = \pi_{k0j} + \epsilon_{kij},
\]

which indicates how to draw a line through the observations for all pupils in a given school, and how individual pupils deviate from this line.

\[
(6) \quad \theta_{kij} = \pi_{k0j} + \pi_{kij} + r_{kij},
\]

which indicates how to draw a line through the observations of each pupil in a given school, and how each observation in time deviates from this line.

Interesting for self-evaluation purposes are the \( \pi_{k0j} \); how good does a school perform compared with other schools? That full use is made of the data can be seen in the way the \( \pi_{k0j} \) are estimated using Empirical Bayes methods (cf. Raudenbush & Bryk, 1986, p. 150):

\[
(7) \quad \pi^*_{k0j} = w_j \pi_{k0j} + (1-w_j) \pi^*_{0j}
\]

The \( w_j \)'s are weights, that are based on the number of observations within schools. The growth-trajectory for school \( j \) is estimated as a weighted sum of the OLS-estimate of this trajectory (\( \pi_{k0j} \)) and the Empirical Bayes estimate of the mean growth trajectory across schools (\( \pi^*_{0j} \)). For instance, the fewer observations there are for school \( j \) the more the Empirical Bayes estimate is based on the average growth trajectory.
SELF-EVALUATION BY STUDYING GROWTH IN ACHIEVEMENT

To illustrate the self-evaluation procedure as outlined above, data were used from a current study on self-evaluation, in which pupil-, class, and schoolperformance are fed back to the schools to study their effectiveness as compared with a control group that does not receive feedback. Data are available on 373 pupils in 22 groups in lower general secondary schools (mavo). Each pupil was tested three times: at the start of grade 1, at the end of grade 1 and at the end of grade 2. The tests were vertically equated and the ability scores for each pupil were calculated using the one parameter logistic model (Verhelst, 1992). For this example K=2, in other words linear growth is assumed. The data used in this example are the data that were fed back to the schools. The assessment of the effects of the experiment will be based on a test to be administrated to the pupils at the end of grade 3, and will be reported by the end of 1994.

Some descriptive statistics of the variables are presented in Table 2.

There are 8 different ways to estimate the performance of a school.

A.1. The most simple statistic is the mean (raw) test score of all pupils within a school at time \( t=2 (\bar{y}_{2,j}) \).

A.2. The next one is based on the mean ability score of the pupils at time \( t=2 (\bar{\theta}_{2,j}) \).

A.3. The third is based on an Empirical Bayes estimate of the score of the pupils after two years of schooling, making full use of prior knowledge on the achievement of the pupils at the two earlier timepoints \( \left( \bar{\theta}_{100} + \bar{\theta}_{000} \right) + 2(\bar{\theta}_{100} + \bar{\theta}_{000}) \), in which \( \bar{\theta}_{100} + \bar{\theta}_{000} \) is the estimated base achievement level for school \( j \) at time \( t=0 \), and \( \bar{\theta}_{100} + \bar{\theta}_{000} \) is the estimated linear rate of change between two consecutive time points.

These first three indicators (A.1. through A.3.) are based on the achievement of pupils after two years of schooling. These indicators can be especially useful when comparing achievement levels with absolute standards, assuming that these are defined on the mathematics ability scale as used in the tests. In this respect indicators A.2. and A.3. are superior to A.1. since they project the scores of the pupils on this scale in a psychometrically correct way. A serious drawback of these indicators A.1. through A.3., however, is that they do not reflect solely the quality of (outputs of) the educational practices, but input differences between schools as well. The next indicator takes some of these input differences into account.

B.1. The fourth estimate of the performance of a school is as A.3. but with a correction for initial differences between pupils with respect to sex and aptitude and with a correction for aptitude effects on growth. This performance index may be conceived as the mean overachievement of the pupils in a school.

Although indicator B.1. does away with the drawbacks of gross effect measures, one might still argue that a more refined and fair measure should take into account input differences between schools with regard to mathematics achievement. This can be achieved by using “progress” indicators.

C.1. The most simple statistic to assess progress is the difference between the mean achievement score at point 2 and point 0 \( (\bar{y}_{2,j} - \bar{y}_{0,j}) \).

C.2. Somewhat more sophisticated is the same statistic, but now based on the ability scores \( (\bar{\theta}_{2,j} - \bar{\theta}_{0,j}) \).

<table>
<thead>
<tr>
<th>variable</th>
<th>mean</th>
<th>s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>mathematics ability at t=0</td>
<td>-.23</td>
<td>.40</td>
</tr>
<tr>
<td>mathematics ability at t=1</td>
<td>.02</td>
<td>.28</td>
</tr>
<tr>
<td>mathematics ability at t=2</td>
<td>.21</td>
<td>.38</td>
</tr>
<tr>
<td>sex (0= boy; 1= girl)</td>
<td>.52</td>
<td></td>
</tr>
<tr>
<td>aptitude</td>
<td>0</td>
<td>7.81</td>
</tr>
</tbody>
</table>
C.3. The Empirical Bayes estimate of the mean growth in a school is the indicator using information at the three timepoints $(2\pi_0 + \tilde{\pi}_{ij})$.  
C.4. The last indicator is once again the Empirical Bayes estimate of the mean growth, but now once again with the same correction as in B.1. (for sex and aptitude effects on initial status and also for aptitude effects on growth).

The hierarchical linear model used in A.3. and C.3. is:

\begin{align*}
\theta_{ij} &= \pi_{ij} + \pi_{ij} + \epsilon_{ij} \\
\pi_{ij} &= \pi_{0ij} + \epsilon_{ij} \\
\pi_{0ij} &= \pi_{00} + \epsilon_{ij} \\
\pi_{00} &= \pi_{000} + \epsilon_{0j} \\
\pi_{000} &= \pi_{000} + \epsilon_{00}
\end{align*}

In estimating B.1. and C.4. sex and aptitude are introduced as predictors of the initial ability score. Moreover aptitude is also introduced to predict differences in growth rates between pupils. Expressions (9) and (10) therefore are altered into:

\begin{align*}
\pi_{0ij} &= \pi_{00j} + \beta_{ij} sex_{ij} + \beta_{ij} IQ_{ij} + \epsilon_{ij} \\
\pi_{ij} &= \pi_{0ij} + \beta_{ij} IQ_{ij} + \epsilon_{ij}
\end{align*}

The results of the analyses are presented in Table 3, where, for simplicity’s sake, the mean is subtracted from the estimates.

Table 3. Estimates of school effects (ranks between brackets), and Spearman correlations (in the last row) between C.4. and the other estimates.

<table>
<thead>
<tr>
<th>group</th>
<th>net growth</th>
<th>gross growth</th>
<th>ability diff.</th>
<th>raw diff.</th>
<th>t=2 net EB</th>
<th>t=2 gross EB</th>
<th>t=2 ability</th>
<th>t=2 raw</th>
<th>corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3021</td>
<td>-.21 (1)</td>
<td>-.21 (1)</td>
<td>-.28 (1)</td>
<td>-.57 (1)</td>
<td>-.06 (8)</td>
<td>-.07 (8)</td>
<td>-.05 (9)</td>
<td>-.10 (10)</td>
<td>1.00</td>
</tr>
<tr>
<td>5042</td>
<td>-.16 (2)</td>
<td>-.15 (2)</td>
<td>-.20 (3)</td>
<td>-.53 (2)</td>
<td>-.11 (4)</td>
<td>-.15 (4)</td>
<td>-.21 (5)</td>
<td>-.49 (4)</td>
<td>1.00</td>
</tr>
<tr>
<td>3022</td>
<td>-.15 (3)</td>
<td>-.14 (3)</td>
<td>-.20 (2)</td>
<td>-.41 (6)</td>
<td>-.05 (9)</td>
<td>-.11 (7)</td>
<td>-.11 (6)</td>
<td>-.31 (6)</td>
<td>1.00</td>
</tr>
<tr>
<td>2013</td>
<td>-.11 (4)</td>
<td>-.11 (4)</td>
<td>-.16 (4)</td>
<td>-.29 (8)</td>
<td>.01 (12)</td>
<td>-.02 (13)</td>
<td>-.02 (12)</td>
<td>-.19 (8)</td>
<td>1.00</td>
</tr>
<tr>
<td>2231</td>
<td>-.11 (5)</td>
<td>-.10 (5)</td>
<td>-.12 (6)</td>
<td>-.33 (7)</td>
<td>-.11 (5)</td>
<td>-.14 (6)</td>
<td>-.09 (7)</td>
<td>-.24 (7)</td>
<td>1.00</td>
</tr>
<tr>
<td>5021</td>
<td>-.10 (6)</td>
<td>-.09 (7)</td>
<td>-.10 (7)</td>
<td>-.26 (9)</td>
<td>-.15 (2)</td>
<td>-.19 (2)</td>
<td>-.23 (3)</td>
<td>-.42 (5)</td>
<td>1.00</td>
</tr>
<tr>
<td>5041</td>
<td>-.10 (7)</td>
<td>-.09 (8)</td>
<td>-.11 (8)</td>
<td>-.26 (9)</td>
<td>-.15 (2)</td>
<td>-.19 (2)</td>
<td>-.23 (3)</td>
<td>-.42 (5)</td>
<td>1.00</td>
</tr>
<tr>
<td>5141</td>
<td>-.09 (8)</td>
<td>-.10 (6)</td>
<td>-.13 (5)</td>
<td>-.23 (11)</td>
<td>-.10 (7)</td>
<td>-.05 (9)</td>
<td>-.07 (8)</td>
<td>-.42 (12)</td>
<td>1.00</td>
</tr>
<tr>
<td>2371</td>
<td>-.04 (9)</td>
<td>-.06 (9)</td>
<td>-.07 (9)</td>
<td>-.34 (18)</td>
<td>.26 (22)</td>
<td>.36 (22)</td>
<td>.46 (22)</td>
<td>11.35 (22)</td>
<td>1.00</td>
</tr>
<tr>
<td>5142</td>
<td>-.04 (10)</td>
<td>-.04 (10)</td>
<td>-.05 (10)</td>
<td>-.26 (10)</td>
<td>-.05 (10)</td>
<td>-.03 (11)</td>
<td>-.05 (10)</td>
<td>-.18 (9)</td>
<td>1.00</td>
</tr>
<tr>
<td>5022</td>
<td>-.03 (11)</td>
<td>-.03 (12)</td>
<td>-.04 (12)</td>
<td>-.41 (5)</td>
<td>.01 (11)</td>
<td>-.03 (11)</td>
<td>-.03 (11)</td>
<td>-.65 (3)</td>
<td>1.00</td>
</tr>
<tr>
<td>2232</td>
<td>-.02 (12)</td>
<td>-.03 (11)</td>
<td>-.04 (11)</td>
<td>-.23 (15)</td>
<td>.01 (13)</td>
<td>-.05 (14)</td>
<td>-.12 (14)</td>
<td>4.41 (18)</td>
<td>1.00</td>
</tr>
<tr>
<td>2012</td>
<td>-.00 (13)</td>
<td>.01 (13)</td>
<td>.01 (13)</td>
<td>-.22 (14)</td>
<td>.15 (20)</td>
<td>.12 (19)</td>
<td>.17 (19)</td>
<td>4.62 (19)</td>
<td>1.00</td>
</tr>
<tr>
<td>5081</td>
<td>.01 (14)</td>
<td>.02 (14)</td>
<td>.02 (14)</td>
<td>.14 (12)</td>
<td>.00 (11)</td>
<td>-.04 (10)</td>
<td>-.05 (11)</td>
<td>.12 (13)</td>
<td>1.00</td>
</tr>
<tr>
<td>2101</td>
<td>.02 (15)</td>
<td>.02 (15)</td>
<td>.04 (15)</td>
<td>-.50 (4)</td>
<td>-.12 (3)</td>
<td>-.14 (5)</td>
<td>-.29 (2)</td>
<td>-.95 (5)</td>
<td>1.00</td>
</tr>
<tr>
<td>2373</td>
<td>.04 (16)</td>
<td>.03 (16)</td>
<td>.04 (16)</td>
<td>.22 (13)</td>
<td>.06 (16)</td>
<td>.09 (17)</td>
<td>.10 (17)</td>
<td>3.05 (15)</td>
<td>1.00</td>
</tr>
<tr>
<td>2401</td>
<td>.08 (17)</td>
<td>.07 (17)</td>
<td>.08 (17)</td>
<td>.52 (20)</td>
<td>.25 (21)</td>
<td>.30 (21)</td>
<td>.40 (21)</td>
<td>10.05 (21)</td>
<td>1.00</td>
</tr>
<tr>
<td>2372</td>
<td>.12 (18)</td>
<td>.11 (18)</td>
<td>.13 (18)</td>
<td>.45 (19)</td>
<td>.13 (19)</td>
<td>.18 (20)</td>
<td>.21 (20)</td>
<td>6.10 (20)</td>
<td>1.00</td>
</tr>
<tr>
<td>5082</td>
<td>.15 (19)</td>
<td>.15 (19)</td>
<td>.20 (19)</td>
<td>.31 (16)</td>
<td>.04 (15)</td>
<td>.07 (15)</td>
<td>.07 (14)</td>
<td>.78 (14)</td>
<td>1.00</td>
</tr>
<tr>
<td>2402</td>
<td>.20 (20)</td>
<td>.20 (20)</td>
<td>.27 (20)</td>
<td>.32 (17)</td>
<td>.03 (14)</td>
<td>.00 (12)</td>
<td>.01 (13)</td>
<td>.66 (11)</td>
<td>1.00</td>
</tr>
<tr>
<td>5111</td>
<td>.25 (21)</td>
<td>.25 (21)</td>
<td>.32 (21)</td>
<td>.60 (21)</td>
<td>.08 (18)</td>
<td>.09 (16)</td>
<td>.10 (16)</td>
<td>3.21 (16)</td>
<td>1.00</td>
</tr>
<tr>
<td>5112</td>
<td>.30 (22)</td>
<td>.29 (22)</td>
<td>.38 (22)</td>
<td>8.57 (22)</td>
<td>.07 (17)</td>
<td>.12 (18)</td>
<td>.09 (15)</td>
<td>4.33 (17)</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Corr. 1.00 1.00 .99 .81 .60 .59 .53 .54
The 22 groups are ordered on the basis of their scores on indicator C.4., being an estimate of the growth in grade 1 and 2, corrected for sex and aptitude (defense mechanisms like “we don’t have very bright pupils” are therefore ruled out). C.4. and C.3. almost lead to the same ranking of the 22 groups.

The ranking is also very similar (.99) with the more convenient indicator C.2., but that is only so because there are only three timepoints in the example! When more timepoints are added the difference in ranking between C.2. on the one hand and C.3. and C.4. will become more manifest. Notice also that C.4. and C.3. have less variation than C.2., which is the direct consequence of the Empirical Bayes approach, that works like “regression to the mean”: the less information there is for a certain school the more its estimate is shrunk toward the grand average.

Greater differences in ranking occur when using the (easy to compute) difference score of the two raw test scores (the correlation now reduces to .81).

Notice, for instance, the misleading information that is presented about group 2011 (at rank 15 at C.4.), that is performing on average according to C.4., but is doing worse on C.1 (at rank 4). The opposite occurs for group 2371; that is at rank 9 at C.4. and at rank 18 at C.1. A complete other ranking occurs when applying B.1. and A.1., A.2., or A.3., as can be readily deduced from the magnitudes of the correlation coefficients (.60, .59, .53, and .54 respectively).

Which of these 8 indicators of school performance should be preferred in case of school self-evaluation? It should be the one that conceptually reflects the effort that the school has made to make its pupils achieve. Stated otherwise, the one indicator that is of primary interest in case of self-evaluation is similar to the preferred criterion in school effectiveness studies, and to the one that one would use in informing parents in a fair way about the school of their children. Our first contention is that one would refrain from using cross-sectional data only, since these may reflect input differences between schools. For school self-evaluation purposes it is of course interesting to find out whether a school meets a certain standard, e.g. a cut-off score on the scale. For this reason A.2. is a better indicator than A.1., the one that is most likely to be used. Chance fluctuations are always possible, and for this reason indicator A.3. is a statistical improvement. The indicators that conceptually come closest to a fair definition of a school effect are the indicators that reflect the progress that pupils make, with, mutatis mutandis, as limitations and advantages the ones just mentioned for the indicators A.1. through A.3. The real problem with these progress indicators is the assessment of the initial achievement level (the base), since in some cases it may not make much sense to test pupils, for instance, for their initial knowledge of Greek language. For this reason one may assess the base level somewhat later than at school entrance, or one may choose to use the overachievement indicator B.1. instead. This is the one proposed by Williams (1992).

Our second contention is that there are a lot of psychometric requirements to be imposed on achievement tests (see also: Seltzer, Frank & Bryk, 1994). Advanced pupil monitoring systems meet these standards. These, together with the conceptual deliberations regarding a fair definition of a school effect, lead to the conclusions that pupil monitoring systems provide a sound basis for school monitoring systems. This brings us to the question of how schools can find those grades and/or teachers within a school that mostly affect the effectiveness of a school in a certain subject area.

Statistically this can be achieved by introducing an intermediate level in the hierarchy between pupils and schools, assuming that the tests are constructed in such a way that, on average, linear growth within and across grades is the case. (This latter assumption can be relaxed by introducing a fixed grade-effect on growth). Comparing each grade in a school with all comparable grades in a sample actually may lead to the identification of poorly operating classes and educational practices. In this way schools learn where to locate the leverage point for change. If, moreover, schools also use an indicator system for assessing these educational practices (e.g. instruments to measure "allocated learning time", or grouping practices), links can be made between the quality indicators of process and output. The procedure to follow in this latter case depends on the characteristics of the self-evaluation system as implemented. First of all, if schools perform below average on the achievement tests as well as on some of the process
indicators, this information can be used to detect those practices where improvement is possible. A more sophisticated approach is, that schools, knowing the production function (in terms of a regression model), decide on the basis of their observed achievement gain, process scores, and the (known) regression coefficients, how improvement can be optimized by implementing change in some of the process characteristics. Lastly, it is possible, that schools seek for internal points of reference (comparing one class with other classes, or making comparisons in time), thus trying to find leverage points for change.

USE AND GENERALIZATION

Process-product relations
When a school would have a full-fledged pupil monitoring system in place, the condition of having primary output data for self-evaluation purposes would be met in a most convincing way.

Measurement of the key-criteria variables, however, is only one vital aspect of evaluation, the other being the internally and externally valid causal attribution of outcomes to antecedent conditions. In other words the natural interest in school self-evaluation will be directed at process-product relationships. Questions like: why do the results in a certain year fall back with respect to the previous year: why do the results with method A appear to be better than with method B, etc.

It is beyond the scope of this article to discuss this issue in any detail, but two lines of approach shall be mentioned that are possible when one deals with the type of output information that comes available from pupil monitoring systems.

First, within a school, teaching conditions could be compared. Particularly when age groups would be in parallel classes (e.g. two first grades, two second grades, etc.), they would form a natural basis to conduct "experiments of nature" (exploiting natural variance) or even planned quasi-experiments. In the latter case Campbell's famous dictum of "reforms as experiments" could be generalized to the within schools level.

In the absence of parallel classes within subject designs as described could be employed to check whether certain innovations or "interruptions" would show up in the achievement patterns.

Secondly, organizational measures at the school level, or school conditions in general, could be similarly assessed by means of within subject designs.

Thirdly, when several schools, for instance within a certain neighbourhood, would employ the same pupil monitoring system and coordinate data collection (i.e. measure at the same levels at the same points in time), all types of causal analysis, with respect to both variables at school and class level, that we know from empirical school effectiveness studies, would be applicable.

In this way networks of schools might start to function as committees of organizations that learn from each other.

Table 4. Prerequisites for output oriented school self-evaluation.

- availability of tests, preferably meeting item-response models
- user friendly approaches to test administration
- facilities (preferably computerized) for data storage, analysis and retrieval
- manpower for internal control and administrative tasks surrounding a monitoring system
- consensus on the use of a monitoring system among staff
- support of school management
- coordinated use of a monitoring system
- proficiency in data analysis, quasi-experimental designs
- meetings of staff and management to discuss reports and to discuss courses of action
Practical prerequisites

School self-evaluation as described in this article has some utopian features. We realize that in actual practice a lot of barriers need to be taken. So, in order to avoid the accusation of being unrealistic, it is perhaps good to summarize the prerequisites for the type of school self-evaluation that has been the center of attention in this article. This is shown in Table 4.

It is quite evident that several of these functions would require the support by external agencies, such as institutes for test development or educational research specialists.

Self-evaluation using product data should take into account the psychometric, statistical, and conceptual deliberations concerning the measurement of school effects. Further developments in computerized testing and in computer software for use in pupil monitoring systems might also lead to the growth of a generation of software for school monitoring systems that minimizes external support.

REFERENCES


Manuscript ontvangen 28-4-1994
Definitieve versie ontvangen 2-9-1994