Grade Repetition or Tailoring the Curriculum?

Paper presented at the Conference on “Cambios Pedagogicos y Fracaso escolar”
[Educational Changes and Failure at School]
Madrid, September 8, 2000

Aletta Grisay

Abstract

The first part of this presentation briefly points out the problem arising, in all school systems, due to the heterogeneity of students' school performance. It attempts to assess the efficiency of grade repetition as a device intended, in theory, to address situations of school failure, by relying, on the one hand, on the results of research where groups of repeaters were compared with control groups with the same difficulties, but who were promoted, and on the other hand, on international surveys enabling the assessment of academic achievement in countries practising grade repetition vs countries that do not practise it.

The second part describes a research action carried out in a sample of 50 schools in the French Community of Belgium for the purpose of reducing the occurrence of grade repetition. Students in these schools were tracked from grade 1 through grade 6. The study, conducted in co-operation with the schools' principals and staff, focused on end-of-year teacher-made exams used to determine whether students should be admitted to the next grade. The analysis of the content of these exams showed that a large number of grade repetition cases were related to deviations from the official curriculum, to the very unequal difficulty of the exams that the students had to sit, and to the biased perception of the heterogeneity of their classes that teachers derived when using their marking system. Refocusing of the curriculum taught and fewer instances of grade repetition were observed upon completion of the intervention.
*Learning difficulties* are a reality that has been with us down through the ages in all places and in all individuals. There exists in every human group, significant diversity linked to aptitudes, and to the greater or lesser speed at which individuals assimilate new knowledge or skills, interests, curiosity, and motivation. We also know very well that even an exceptionally gifted person may know nothing about music, or may be an inept swimmer, a boring speaker, an execrable cook, a bungling handyman, etc.

There is nothing universal, however, about *school failure*. That is an historically recent phenomenon, linked to the emergence within our societies of an institution charged with dispensing *certain types of instruction* to a *specific age group* (children and adolescents) according to *codified methods* (standards for premises, timetable, content and how it is divided, students’ ages, class organisation, assessment and certification, etc.). A child who fails is a child who the school institution judges, at a given moment, to have not managed to learn what it estimates reasonable that children of his/her age should have learned. The institution therefore deems necessary to do something to correct the situation. By this, we see that the school failure situation cannot be summarised only to the student’s learning difficulties. Failure is also, to some extent, the official recognition of these difficulties. *School failure is often what the school says about or does to a student with problems.*

Most “traditional” decisions that our educational systems apply in such cases are based on the idea that if a student is experiencing difficulties, it must be that his/her proficiency level is too different from that of other students in his/her class to allow him/her to benefit from the teaching that is offered. It is therefore necessary to reduce this heterogeneity by placing the student in an environment that will better suit him/her and in which the requirements are more to his/her measure: a stream with a less challenging curriculum, or an ability class or ability group composed of students with similar difficulties, or by having him/her repeat the school year that he/she has just completed.

It is the latter of these decisions, grade repetition, to which this article will be devoted. Drawing on results from a large number of studies that have endeavoured to assess its effectiveness, we shall show that there are very few chances of achieving, by this means, a real reduction in the heterogeneity of achievement or of improving the situation of students in difficulty. We shall next analyse the specific case of an educational system in which grade repetitions are particularly numerous (that of the French Community of Belgium) to examine in greater detail the various negative side effects engendered by this practice and outline alternative schemes1.

**Is grade repetition of any benefit to weaker students?**

In countries where grade repetition is accepted, this practice is usually the subject of a sort of tacit consensus among teachers and students’ parents, according to which *it is better to have a weak student repeat a grade than to let him/her pursue a curriculum where his/her difficulties will only be aggravated.*

---

1 A more comprehensive review of the international literature on grade repetition and of research on the case of the French Community of Belgium can be found in M. CRAHAY, *Peut-on lutter contre l’échec scolaire ?* [Can school failure be fought?], Brussels, De Boeck, 1996, from which the present article draws inspiration.
In theory, this is a defensible point of view. A repeater is often a slow student, so it would seem appropriate to allow him/her more learning time. He/she is also sometimes a somewhat immature student, whose behaviour, it is hoped, will improve with age. But in practice? If there is a field in which empirical research results are without ambiguity, then it is this one. As a remedial measure, grade repetition can only be formally decreed inadvisable, since its effects have been proven to be either harmful or (in the best case) ineffectual, and in such a general way, that no reasonable doubt should subsist.

The author of the most recent meta-analysis of the effectiveness of grade repetition (Holmes, 1990) drew on 63 studies in which the progress of groups of repeating students was compared to those of students with the same initial characteristics (in terms of standard test scores) but who were allowed to move on to the next grade. The studies examined a variety of criteria: cognitive progress (gains achieved, after one or several years, in achievement tests, or marks given by teachers) but also attitudes toward school, absenteeism, social adjustment, emotional well-being, etc.

The results, presented in Table 1 below, are expressed in terms of size effect, i.e. the average difference in achievement scores observed across all available studies between repeating and non-repeating students, divided by the standard deviation of scores for the group of those promoted.

The average scores obtained by repeaters prove to be significantly lower than those of promoted students of the same initial level for all criteria relating to academic achievement. It is interesting to note that the greatest variance concerns classroom tests, used by the teachers themselves evaluate these students. This seems to indicate not only that achievement among the repeaters becomes objectively less satisfactory, but that, in addition, it tends to be underestimated by their teachers.

Progress is also negative for practically all affective criteria, although less markedly so. Oddly, scores are very slightly positive for the students’ self-image. This is probably attributable to an artefact: scores on the self-image scale for students in difficulty tend to be more often affected than those of other students by biases attributable to the social desirability of answers.
Table 1: Differences observed between initially comparable groups of repeating students and those who were promoted

<table>
<thead>
<tr>
<th>Measured criteria</th>
<th>Number of studies</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global academic proficiency</td>
<td>47</td>
<td>-0.31</td>
</tr>
<tr>
<td>Mother tongue</td>
<td>18</td>
<td>-0.33</td>
</tr>
<tr>
<td>Reading</td>
<td>34</td>
<td>-0.30</td>
</tr>
<tr>
<td>Mathematics</td>
<td>31</td>
<td>-0.25</td>
</tr>
<tr>
<td>Science</td>
<td>3</td>
<td>-0.37</td>
</tr>
<tr>
<td>Teachers’ made tests scores</td>
<td>3</td>
<td>-0.78</td>
</tr>
<tr>
<td>General affective reactions</td>
<td>27</td>
<td>-0.21</td>
</tr>
<tr>
<td>Social adjustment</td>
<td>27</td>
<td>-0.21</td>
</tr>
<tr>
<td>Emotional well-being</td>
<td>10</td>
<td>-0.12</td>
</tr>
<tr>
<td>Behavioural attitudes</td>
<td>10</td>
<td>-0.23</td>
</tr>
<tr>
<td>Self-image</td>
<td>11</td>
<td>+0.06</td>
</tr>
<tr>
<td>Interest in school</td>
<td>10</td>
<td>-0.18</td>
</tr>
<tr>
<td>School Attendance</td>
<td>5</td>
<td>-0.22</td>
</tr>
<tr>
<td>Global size effect</td>
<td>63</td>
<td>-0.26</td>
</tr>
</tbody>
</table>

Source: HOLMES, 1990 (adapted by CRAHAY, 1996)²

² The studies used by Holmes are of Anglo-Saxon origin. Those conducted by French-speaking research teams are rarer (SEIBEL, 1984; BAIN, 1988). All nevertheless confirm the trends presented in Table 1.
The question remains as to why such a clearly ineffective “therapy” is still used in many educational systems and from whence comes the conviction of so many teachers that repeating is beneficial, if not for all, then at least for many students. Teachers are neither irrational beings nor insensitive automatons. Their perceptions of what goes on in their classes and what may help or hinder their students are often keen and correct. Why is it then that their assessment would fail in this case? Some of the data collected by Holmes (1990) can be used to propose an explanation.

In some of the studies analysed by the author, data for repeaters and those promoted were collected several times in the years following repeating of a grade for some and promotion for others. This permits two types of comparisons:

- same grade comparison (performance for the two groups is assessed for the *same grade*, which repeaters attain, of course, when they are *one year older* than those promoted);

- same age comparison (performance for the two groups is assessed at the *same age*, although they are in *different grades*).

These comparisons are presented in Tables 2 and 3.

**Table 2: Studies based on a same grade comparison between repeaters and promoted students**

<table>
<thead>
<tr>
<th></th>
<th>1 year later</th>
<th>2 years later</th>
<th>3 years later</th>
<th>More than 3 years later</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of studies</strong></td>
<td>10</td>
<td>7</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td><strong>Size effect</strong></td>
<td>0</td>
<td>+ 0.02</td>
<td>- 0.12</td>
<td>+ 0.04</td>
</tr>
</tbody>
</table>

Source: Holmes, 1990 (adapted by Crahay, 1996)

This first comparison shows that for the same grade, the repeaters’ handicap is, in fact, barely visible. Their scores do not differ significantly from the ones achieved by those promoted a year earlier upon completion of the same grade. Strictly speaking, the repeated year is thus not unproductive. Repeating students do improve somewhat, which gives the teacher who retained the repeaters in his/her class the impression that the measure taken was useful. What the teacher cannot verify is that the repeated year is *significantly less productive* than it would have been if these same students had been promoted. The relevant comparison, which the teacher is never able to make, is the same age comparison, which appears in Table 3. It shows that those promoted progressed much faster than the repeaters, and that the difference only widens over the years.
Table 3: Studies based on same age comparisons of repeaters and promoted students

<table>
<thead>
<tr>
<th></th>
<th>1 year later</th>
<th>2 years later</th>
<th>3 years later</th>
<th>More than 3 years later</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of studies</strong></td>
<td>28</td>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td><strong>Size effect</strong></td>
<td>-0.41</td>
<td>-0.64</td>
<td>-0.74</td>
<td>-0.88</td>
</tr>
</tbody>
</table>

Source: HOLMES, 1990 (adapted by CRAHAY, 1996)

These results are probably explained by the fact that repeaters and those promoted were not exposed to the same curriculum. The repeaters merely returned to material already tackled, whereas those promoted were “pulled upward” by the pressure of a more challenging curriculum. The year both groups spent in their respective environments was unequally stimulating, and the available time unequally exploited.

***

Let us also point out that Holmes’ meta-analysis is only the most recent in a long series and that it just confirms facts that have been known since the earliest studies on repetition, as far back as the beginning of the 20th century. Other researchers have investigated on various collateral aspects of repetition, and the results are not very encouraging:

1. When grade repetition is accompanied by individual monitoring of students, then results are somewhat less negative (Peterson et al., 1987). However, the repeaters being individually monitored do not achieve higher scores than those of their classmates with the same initial level who were promoted.

2. Grade repetition is not always referred to by its name, but the “indirect” forms that it sometimes takes are shown to be just as inefficient. For example, it also occurs in the case where admission to primary school is delayed for one year for children judged to be immature or at an insufficient level to adjust to formal education. According to available research (Gredler, 1984, Shepard, 1990), the extra year of pre-primary education offered to such children – in the form of repeating the last year of nursery school or enrolment in a transition class – is either ineffectual or harmful.

3. Along the same lines, we should have reservations about reforms adjusting the length of study cycles\(^3\) (where, for example, students in difficulty are permitted to cover in three years a cycle of studies that other students cover in two years), since this does not produce very convincing results. Nevertheless, a survey conducted by the French Ministry of Education on an optional “3-year cycle” at the start of secondary education (Ernst & al., 1994) concluded that students who had been given this opportunity were in part able to make up for their lag. Their score at the end of the cycle was slightly higher (+0.158) than that attained by students with the same initial competencies and who completed the cycle in 2 years. Unfortunately, the authors compared scores only for the same grade; as we

---

\(^3\) The periods of study in which education is subdivided in many educational systems.
have seen, this comparison is fallacious since it does not consider the faster progress of the promoted students during the extra year that their classmates spent in the same cycle.

4. Last, several studies show that for the same achievement level, students whose school career includes a grade repetition are significantly more likely to drop out of school. A path analysis presented by Grissom and Shepard (1990), based on various very large data-bases (several tens of thousands students), indicates that grade repetition has a specific influence on the probability of dropping out, which is partly independent of effects attributable to the students’ learning difficulties. According to the authors, the role played by this factor in the decision to drop out of school would be even more determining, in many cases, than students’ actual achievement levels.
Are educational systems that practice grade repetition more effective than those that do not?

A number of countries have adopted a system of automatic promotion of students during all or part of compulsory education. Among industrialised countries, these include the Nordic countries (Denmark, Finland, Norway and Sweden), the United Kingdom, Ireland, and Japan. Among countries that resort to grade repetition, some do so only on an exception basis (e.g. Italy, Greece, and Germany) whereas in others the proportion of students who have had to repeat at least one grade during their schooling career can be quite significant (e.g. France, Belgium, and Portugal). What can be said about the effects of these policies, in terms of overall academic achievement, of students’ heterogeneity, and of access to diplomas?

Academic achievement

In surveys comparing the performance of educational systems conducted by the IEA (International Association for the Evaluation of Educational Achievement), students in countries that practice grade repetition should, in principle, score higher than those in countries that seldom do so or not at all. This is because in each participating country the same test is administered to the modal grade attended by the majority of students of a given age cohort (e.g. 9-year-olds or 14-year-olds). Therefore, countries in which the weakest part of the tested age cohort group lags behind (and therefore is not yet in the modal grade) have an advantage. Their sample does not include students who are lagging behind, or if it does, they are at least one year older than other students attending the grade in which the test is administered.

However, this hypothesis is not borne out. In most of the available international studies (e.g. the IEA Reading Literacy Study (1991), or the IEA Third Mathematics and Science Study (1995), or the OCDE International Adult Literacy Study (1995)) tests scores do not appear to be lower for those countries in which promotion is automatic. In fact, the opposite is often true. Japan, for example, often tops the list for mathematics and science, and Finland, Sweden, and Norway score significantly higher than the international average for reading literacy. Countries that make the heaviest use of grade repetition exhibit several weak points. France, for example, scores honourably in reading literacy and mathematics but mediocre in science. Belgium is a significant case. In the French-speaking part of the country (in which repetition rates are very high), science and reading literacy scores are among the lowest in industrialised countries. In contrast, the Flemish-speaking part (which resorts much less to grade repetition) scores significantly higher than the international mean in mathematics and science.

Heterogeneity

One of the most surprising findings from these surveys is the low impact of the practice of grade repetition on student population heterogeneity as regards academic achievement. The 1995 IEA/TIMSS survey, for instance, showed that the difference in mathematics and science test scores between the 25% weakest and the 25% strongest students in each country is very high. It represents 1.20 times the standard deviation of scores distribution. In order to fully appreciate the amplitude of this heterogeneity, one should know that it is equivalent to almost four times the difference in achievement observed between two adjacent grades for mathematics, and more than three times the same difference for science. In other terms, the
weakest quarter of the population accumulated a lag of four school years in mathematics and three years in science versus the strongest quarter. The interquartile range is of the same order of magnitude for reading literacy achievement measured in the 1991 IEA/RLS survey.

In these conditions, making even a significant proportion of students repeat one grade cannot lead to a significantly higher homogeneity in the proficiency level of the population enrolled in the same grade. Actually, in some countries that resort intensively to grade repetition, the interquartile range in achievement scores is narrower than the international average (0.85 standard deviation in Portugal, 0.107 in France for mathematics). In other countries, however, like Belgium and the Netherlands, the interquartile range is wider than the range observed in some of the countries in which promotion is automatic, e.g. Sweden, Norway, and Denmark. It can be said, in this last case, that grade repetition does not attain even its least ambitious goal, i.e. making the work of schools easier by reducing disparity between students attending the same grade, through delaying by one year the promotion of the weaker students.

**Access to the secondary school leaving diploma and to post-secondary education**

The OECD regularly publishes an estimate of the rate of secondary school diploma attainment in the industrialised countries, i.e. the number of diplomas that are awarded in a given year, divided by the number of young people of an age to obtain the diploma. In 1998, the average ratio in the 27 industrialised countries for which data were available was 85%. Dispersion was relatively low: only one country (Mexico) had a ratio lower than 70%; the ratio was in the 70-80% range in five countries, in the 80-90% range in eleven countries; it exceeded 90% in seven countries – three of which practice automatic promotion (Finland, Norway and Japan). Ireland is the only country practicing automatic promotion that shows a significantly lower graduation ratio (79%), compared to the international mean in industrialised countries.

It should be furthermore noted that the indicator gives a slight advantage to countries that practice grade repetition, inasmuch as the number of diplomas awarded in a given year includes not only diplomas awarded to students completing secondary school at the normal age, but also those who were lagging behind by one or more years. Manifestly, using a system of automatic promotion does not entail a handicap in terms of numbers of diplomas awarded.

On the other hand, countries that have adopted a system of automatic promotion often show enrolment rates in post-secondary education that are higher than the average for industrialised countries. In 1998, these rates were higher than 50% in Finland, Denmark, Sweden, and Japan, versus an average of 41% for the industrialised countries.

There is thus no statistical confirmation for the notion according to which educational systems that have done away with grade repetition are at risk of a levelling-down of academic achievement.

**Improving the monitoring of curricula and the assessment of proficiency in attaining them: an alternative to grade repetition?**

The case of the Belgian French Community’s educational system is of particular interest for the subject under discussion. Grade repetition is a particularly frequent occurrence in that system, where almost one quarter of students do not complete primary schooling without repeating at least one grade, and less than 40% complete secondary schooling “on time”. In
terms of educational costs this means, approximately, that the salary of one teacher out of nine is spent in dealing with grade-repeating students.

In light of international comparisons, educational achievement proves to be abnormally weak in several areas, notably reading literacy and science, whereas Flemish-speaking schools in the same country – which practice grade repetition to a much lesser extent – show much better scores in mathematics and science. During the last 15 years, the Service de Pédagogie expérimentale of the University of Liège devoted several studies to the examination of practices used in French-speaking schools to decide whether students should be promoted. These studies bring to light various mechanisms that in our view help explain these negative phenomena and point the way to alternative solutions.

First, a few words on the context. Belgium is a federal state in which the educational systems of the three linguistic Communities (Flemish, French and German-speaking) enjoy a high degree of autonomy. In each of the Communities, various organising bodies are in charge of running schools. There are three competing types of schools: schools organised by each Community (under the direct authority of the Ministry of Education of each linguistic region), schools organised by the local public authorities (cities, municipalities and regions) and an important network of denominational schools (mostly Catholic) which are private, but funded by the government.

There is no national system of certification of students’ achievement at any point during the schooling career. Promotion and retention decisions are made at the end of each school year by the school’s staff, mostly based on teacher-made tests. Test scores may be combined with (or replaced by) an assessment of students’ work throughout the school year. Each school authority develops their own curricula (subject to approval by the Ministry) and supervises their implementation through its own inspectors. A central body of inspectors under the authority of the Minister is in charge of verifying compliance with a number of common regulations pertaining to security, size of classes, teachers’ qualifications, etc. At least in principle, this body is also in charge of checking that curriculum standards do not vary too widely from one school to the next.

Precisely at the behest of two of these inspectors, the University of Liège undertook the first of a series of studies on grade repetition (Grisay, 1984), the results of which are summarised here.

These inspectors had noticed that repetition rates in primary schools enrolling students from a privileged socio-economic background, whose achievement levels appeared to be quite satisfactory, were just as high (or at times even higher) than those in schools attended by problem-ridden underprivileged populations. They were wondering whether tests taken by students in these different types of schools were not of uneven difficulty.

An external mother tongue (French) test was prepared and administered to all students in the last grade of primary school in the two inspectors’ districts. Test scores were compared to the marks obtained by the same students in their end-of-year, teacher-made French examination. The analysis confirmed the inspectors’ hypothesis. Within each class, correlations between examination marks and test scores were indeed high (almost all fall within the .70-.80 range, with a median value of .75), indicating that rankings of students based on external test scores and teacher-made exam scores were quite similar. On the other hand, correlation across the
entire student population was only .52, which means that from one class to the next, examination marks do not have the same “value” in terms of students’ proficiency measured by the external test.

The relationship between examination marks and external test scores was thus shown to be complex (and, in technical terms, suppressive), due to the fact that it was positive within classes but negative between classes. The examinations devised by teachers tended to be more difficult in classes where the proficiency level (as measured by the test) was higher than the average, whereas they were easier in the “weaker” classes. The result was that for the same proficiency level, students experiencing difficulties were more likely to have to repeat a grade if they were enrolled in a stronger rather than a weaker class.

On the other hand, and precisely because of the uneven difficulty of examinations, the marks awarded by teachers tended to overestimate, in a way, the heterogeneity of classes. By comparison to the test scores, the best students in each class tended to receive marks that were proportionally higher than “deserved”, whereas the weakest students tended to be “underestimated”.

Figure 4 illustrates this phenomenon through three concrete examples, showing, for three of the sampled classes, the relations between students’ French examination marks and their test scores. To make the graph easier to read, the two results (expressed as standard scores) are represented on parallel scales and linked by a line – each line thus corresponds to the results for one student. Circled letters correspond to students who, based on their exam score, were required to repeat the grade. Only about thirty students are represented, i.e. not all students in the three classes.

Class I is a particularly weak class. With respect to the entire population, the average test score is \(-1.6\) standard deviation (\(\sigma\)). The distance between student A (the weakest in the group) and student G (the strongest) is about one standard deviation in terms of test scores. The teacher’s marks stretch out this distribution significantly. Student A receives approximately the mark he/she “deserves” (with respect to the overall population), but all other students are “overestimated”, with the result that the gap between A and G (promoted to the rank of “good student” by a mark higher than \(+1.4\ \sigma\) ) more than triples.

In class II, test scores for the majority of students are clustered around the mean. The distance between student A and I (respectively the weakest and the strongest in the group) is \(1.5\ \sigma\). The marks awarded by the teacher reflect a sharper contrast: the weaker students tend to be “underestimated”, whereas the best students of the class are “overestimated”. The distance between students A and I attains \(4\ \sigma\).

Class III is a strong class, in which many students obtain excellent test scores. As can be seen, weak students are not the only ones to be “underestimated”: that fate also befalls the “less strong” of the group of strong students, leading to the amplification of distances between students in that group. Students E and F (who score significantly higher than the mean for the entire population in the test) thus become “poor students” and will be made to repeat the grade, although their test scores are scarcely lower than those of their classmates G, H and I – considered “good” students by the teacher.
In all three examples, the same mechanism of adaptation of school marks to the performance level of the class is at work. In our view, the examples underscore the highly relative character of a notion such as heterogeneity, on which – in principle – grade repetition is founded. A homogenous class may appear to be disparate to the teacher, because the examination and the proficiency scale he/she uses are adapted to the level of the students. Taking the example of class I, the instrument used by the teacher will not register the common weakness on which the group’s homogeneity is founded; on the other hand, it will be sensitive to finer variations among students than those registered by the external test.

Hence, the aggregation of students into classes tends to underscore the existing differences between individuals comprising the population (at the risk, perhaps, of further accentuating them...).

Whereas differences between students tend to be overestimated, differences between classes are instead significantly underestimated. In the examination, the average mark for class I is very close to that of class III, whereas the latter performed much better in the test.

---

4 For this reason, it would be incorrect to consider that the teacher’s marking is subjective. The differences it depicts are real enough, but their importance is somewhat amplified by the absence of a reference to a wider reality than the class. We prefer to refer to the relative nature of such marking.
Variations of proficiency between classes are thus “masked”: that is also a direct result of the fact that examinations are adapted to the level of each class. Actually, the assessment performed in each class is quite simply independent of external variations.

One of the important consequences of this phenomenon is that remedial measures are based only on proficiency variations between individuals and never or rarely on collective weaknesses. Obviously, all students in class I will never be made to repeat their grade, although each one of them is much less proficient than many candidates for grade repetition in the entire sample. The teacher in charge of class I and the school headmaster, even if they know that it is a weak class, may well be unaware of the seriousness of the situation. In this instance, it seems clear that cases like class I call for urgent and drastic intervention, to avoid the catastrophe awaiting all those students upon admission to secondary education. Moreover, in the face of such uniformity in failure, it could well be that the quality of teaching was wanting during the year or during previous years, hinting at the probability that the school itself needs remedial measures.

The magnitude of the phenomenon of overestimation of differences between individuals was estimated: it explains 15 to 20% of the variance of examination marks, which is considerable. It entails, in particular, an aggravation of inequalities related to students’ gender, age, socio-economic and cultural origin. On average, underprivileged students are in fact less proficient than their classmates. They are thus more often at the bottom end of the ranking, where they are moreover exposed to a higher risk of underestimation. The risk is not negligible: about one half of the previously discussed scoring bias related to overestimation of student heterogeneity (accounting for 8% of the variance of examination marks) concerns these students. Let us be clear that in no way are we referring to the effects of prejudice that the teachers could possibly entertain vis-à-vis underprivileged students. The underestimation of low SES students’ proficiency, in this case, is simply due to the uneven difficulty of teacher-made tests, which itself is the result of the perfectly legitimate desire of teachers to adapt their instruction to the level of the class in their charge.

However, to the extent that refusing promotion to a student has consequences that go beyond the narrow confines of the class, these disparities are hardly acceptable at the primary level, in which all students are in principle exposed to the same curriculum. In fact, scoring biases would be minimal if the average student proficiency levels were, in practice, similar from one school to another and from one class to another. This is unfortunately not the case. The study, which on this point confirms a characteristic of the French-speaking Belgian school system already identified on several occasions by IEA surveys, shows that between classes disparities are particularly significant. In the sample used for the research described here, more than 40% of the total variance in test scores is between-classes variance. It should be noted that between-class variations in French exam marks only represent 17% of total variance – in addition, they only coincide to a small extent with the variations among classes measured by the test.

One of the important conclusions of the study is that in a strongly decentralized school system, where tests on which class promotions are based are left entirely to the teachers’ initiative, the disparities between individual schools tend, to some extent, to disappear (seemingly) and to be transformed (actually) into disparities between individual students, due to the class-level test adaptation mechanism described above. This transformation into inter-individual variance ultimately affects the students when they cross the partitions of the school
system. Student A in class I cited in the example above, when he enrolls in a secondary school, will not only bring with him the weaknesses that differentiate him from the other students in the previous class. He will also bring with him the shortcomings that characterize all of his classmates and who differentiate them from other classes or other schools where the curriculum was either more demanding or better taught. In the new school environment where the student will be placed, these shortcomings will form, henceforth, an integral part of his individual weakness. Therein probably lies one explanation for the particularly high failure rates observed at the various transition points of the educational system (e.g. between primary and secondary education teaching, between the cycles of secondary education or between secondary and tertiary education), as well as the considerable failure risks often associated with a change from one school to another.

***

Following these worrying reports, a thorough review of summative evaluation practices used in French-speaking Belgium was judged necessary by the authorities. This led to the launch of a second study (Grisay, 1988 and 1991), comprising a quantitative phase similar to the previous study (analysis of the grading system used in a sample of classes), but also a more qualitative phase (analysis of exam content and difficulty). The study took the form of action research: the objective was not only to better understand the phenomena concerned, but also to try to reduce repetition rates through better monitoring of teachers' requirements.

**Analysis of exam marks**

Students in fifty Belgian French Community schools were tracked from their entry in grade 1 until their advancement to secondary school. Each year, external mother tongue and mathematics tests, designed in collaboration with the Inspectors and representative, as accurately as possible, of official curriculum requirements, were administered to the students in these classes. The scores, compared to marks obtained on teacher-authored classroom tests, were regularly communicated to each teacher participating in the study in the form of a personalized and confidential file enabling him/her to compare his/her class and its assessment practices to his/her colleagues' average marks and practices:

- Were his/her class test scores higher than, comparable to, or lower than the other classes in the sample? With consideration given to the socio-economic characteristics of the population that attends his/her school, were these scores “expected”, “higher than expected”, or “lower than expected”?

- Did the marks that he/she gave to his/her students suggest that his/her examination was more difficult, easier or of a level of difficulty comparable to the average of the examinations passed by students attending the same grade in the other classes?

- What were the scores obtained on the external test by students to whom he/she refused promotion to the higher grade? Among the remainder of students in the sample who obtained similar scores, how many were also required to repeat the class, and how many were promoted?

These analyses confirmed that the trends already revealed by the preceding study in the last year of primary education teaching were also observed in the other years. The
disparity of the students’ level from one class to another was significant, but only mildly “visible”, insofar as the exams’ difficulty varied according to this level. Marks attained by students tended, on the other hand, to overestimate the heterogeneity of each one of the class groups.

Among the results that were the most upsetting to teachers participating in the study were those that related to the cases of repetition, which often appeared as somewhat arbitrary when information on the student’s proficiency measured by the external test was available. The example shown in table 5 shows the standardized test scores of some 2,000 students tested at the end of grade 2, of which 50 were refused promotion to the higher grade following their exams.

Admittedly, almost all the students who had to repeat obtained test scores that placed them below the population mean. However, many more students in the population, who obtained scores comparable or lower than theirs, were not required to repeat. For example, ten repeaters had standard scores comprised between -1 and -0.5, but that group also included 106 students who were promoted. And 94 other students were also promoted, despite test scores substantially lower than theirs.

Table 5. An example: Scores obtained on an external test by 2nd primary grade repeaters in the sampled schools

<table>
<thead>
<tr>
<th>Among the students who obtained a global external test score* of:</th>
<th>The number who advanced to the higher grade was:</th>
<th>The number who repeated the grade was:</th>
</tr>
</thead>
<tbody>
<tr>
<td>less than -3.000</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>between -2.000 and -3.000</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>between -1.000 and -2.000</td>
<td>81</td>
<td>17</td>
</tr>
<tr>
<td>between -1.000 and -0.500</td>
<td>106</td>
<td>10</td>
</tr>
<tr>
<td>between -0.500 and 0.000</td>
<td>221</td>
<td>0</td>
</tr>
<tr>
<td>greater than 0.000</td>
<td>430</td>
<td>3</td>
</tr>
</tbody>
</table>

* average of standard scores obtained on French and math external tests.

Study of the content of exams

The “qualitative” phase of the study, based on the collection and analysis of the exams papers used in each sampled class, provided results as surprising as the preceding ones.

On the one hand, as one might expect, the summative evaluation practices found in the schools in the sample were characterized by extreme diversity:

- Certain schools organised exams every year, others did so only at the end of each of three two-year cycles of Belgian primary education, others did not have any at all during the first two years;

- In certain schools, a same exam was used in all classes of a same grade; while in others, each teacher developed his own exam;
- The exams also differed both in terms of the points allotted to each domain evaluated and in terms of the weight given to work throughout the year compared to the results from the final assessment;

- The duration of exams, as well as the number and length of exercises assigned varied substantially;

- The questions asked, their instructions, and their formatting were generally clear and properly designed. However, exam papers that were hardly legible, or contained confounding stems, and even content mistakes were not exceptional;

- Most importantly, the nature of the topics evaluated (and the competence level required of the students for each one of them) was extremely variable.

Only three constants emerged:

- **Pointillism:**
  The examinations, essentially, consisted of innumerable drills, which measured the command of knowledge in “smithereens”. More global tasks, assessing the student’s ability to use articulated or integrated strategies were unfrequent.

  An interesting example was provided, in French, by the items evaluating the mastery of the notion of singular and plural, whose evolution could be followed from beginning to end of primary schooling. More than 2,000 items were found in exam questions bearing on this notion. Approximately 70% of them required the student “to put in the plural” an isolated noun phrase where, generally, at least one of the words had an irregular plural (50 to 60% of the cases). The memorizing of the morphological exceptions was thus favoured at the expense of teaching the rules of syntax that govern the grammatical number. Only in the last grade of primary school a majority of items were based on complete sentences - in which, however, morphological exceptions shared the stage with trap problems relating to the rarest or most complex cases of grammatical agreement. It should be noted that at the end of primary education, the items covering the grammatical number, which appeared in the external test (and which included only regular sentences without any exception) were answered correctly by only 40% of the students.

- **Formalism:**
  A clear prevalence was observed, in French as well as in Mathematics, of questions relating to formal knowledge or to the application of stereotypical procedures, rather than to those that measure real comprehension and functional use of the notion learned.

  In French, for example, of 5,000 or so exercises that comprised the curriculum covered over six years, nearly 88% related to formal concepts (primarily grammar and spelling), while reading comprehension and written expression only represented, respectively, 5% and 6% of the content of exam papers. Listening comprehension, for its part, was hardly assessed at all (0,3% of the exercises). A number of cases were observed where a French exam was comprised of only grammar and spelling questions, without any reading comprehension task.

- **Upward drift:**
  All or almost all exam papers (for all grades) contained at least some questions either not required by the curriculum to be taught in that specific grade, or deliberately more difficult than required.
When analysing a sample of math exams passed in grade 4, for example, it was found that in the most “difficult” exam, nearly 70% of the questions referred to topics which should not be have been taught, according to the curriculum, until the next primary education cycle. In the “easiest” exam, “only” 17% were found.

This latter finding was particularly important, for it enabled us to better understand the source of the unequal difficulty of the exams highlighted by the previous study. This was not related to the fact that some of the exams might have been less demanding than required. Actually, all of them exceeded the normal requirements of the curriculum, but some more that others.

The way in which school curricula are designed in the French Community of Belgium was probably responsible, at least partly, for this drift in difficulty:

- With the view of allowing a flexible monitoring of the educational schedule, the objectives to be attained are drawn up, in French-speaking Belgium, for each of several two-year cycles. The curriculum indicates the knowledge and skills, which should be acquired at the end of the second year of each cycle. It does not specify, however, the progression to be followed (topics to be taught in the first year of the cycle or reserved for the second year). In the absence of such an indication, most teachers tend to cover the whole two-year curriculum in the first year.

- A number of curriculum topics relate to pre-awareness objectives, which are intended by the designers to lay the foundation for later instruction (to “initiate” a given notion, to “build awareness” of another, to “impregnate an idea”, to “progressively build notions”, etc.). This has led many a primary school teacher to teach the topics in due form since the beginning, thus anticipating, sometimes considerably, the moment when they should normally be taught. It is difficult to ascertain the proper application of “impregnation”, “sensitising”, or “awareness-building”, and therefore the teacher concludes that it is safer to ‘teach it all’.

- The limits of the subject matter to be taught in each domain are seldom specified ("operate only on whole numbers", “without introducing terminology”, “while limiting oneself to simple cases”). Too often, the curriculum leaves it to the teacher to infer these limits indirectly – from a description of the content to be taught during the next cycle.

- Hardly any distinction is made between basic competencies and activities that can provide possible enrichment. For many teachers, going beyond the curriculum is justified by the need to offer to fastest students, or to a particularly good class, stimulation in line with their aptitudes. Official directives do not really provide them with the means to translate this legitimate concern into controlled differentiation, and not into a headlong rush thereby depriving the weakest students of the time necessary for setting the foundations of their learning.

- Lastly, the official curriculum does not comprise (or has very few) examples making it possible to concretely illustrate the nature and level of difficulty of the tasks that the student should be able to successfully confront at various points during primary education.
The intervention

The research team organized information feedback to the schools taking part in the study. The exam questions in French and Mathematics collected in the 100 classes in the sample were sorted according to the objective of the curriculum that they were assessing. Once tabulated, these data made it possible to visualise, for each objective (and in a very concrete way), what was required, for a same grade, in the various classes in the sample.

These compilations were circulated to the teachers. Each teacher was asked to examine the questions that both he/she and all of his/her other colleagues asked their students to answer for each objective of the curriculum, and to indicate which of them, in his/her opinion, were "well adapted to the students’ grade”, “too difficult for that grade”, or “too easy for that grade”.

The results of this consultation proved to be interesting. While the exam questions that they developed and administered to their students differed considerably in terms of their difficulty, the teachers, on the other hand, were remarkably consistent when they were asked to judge this same difficulty (Table 6).

Table 6: Average percentage of exam questions judged by the teachers to be too difficult, or adapted to the grade concerned, or too easy.

<table>
<thead>
<tr>
<th>MATHEMATICS</th>
<th>Grade 1</th>
<th>Grade 2</th>
<th>Grade 3</th>
<th>Grade 4</th>
<th>Grade 5</th>
<th>Grade 6</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>N = 288 teachers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Much too difficult</td>
<td>13.5 %</td>
<td>12.9 %</td>
<td>10.0 %</td>
<td>11.7 %</td>
<td>10.6 %</td>
<td>9 %</td>
<td>11.3 %</td>
</tr>
<tr>
<td>B. Somewhat too difficult</td>
<td>23.2 %</td>
<td>33.6 %</td>
<td>23.3 %</td>
<td>21.6 %</td>
<td>25.3 %</td>
<td>30.1 %</td>
<td>26.2 %</td>
</tr>
<tr>
<td>C. Well adapted to the grade</td>
<td>61.3 %</td>
<td>52.3 %</td>
<td>64.5 %</td>
<td>64.1 %</td>
<td>59.5 %</td>
<td>55.6 %</td>
<td>59.5 %</td>
</tr>
<tr>
<td>D. Somewhat too easy, or too easy</td>
<td>2.0 %</td>
<td>1.2 %</td>
<td>2.2 %</td>
<td>2.5 %</td>
<td>4.6 %</td>
<td>5.3 %</td>
<td>2.9 %</td>
</tr>
<tr>
<td>TOTAL</td>
<td>100 %</td>
<td>100 %</td>
<td>100 %</td>
<td>100 %</td>
<td>100 %</td>
<td>100 %</td>
<td>100 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FRENCH</th>
<th>Grade 1</th>
<th>Grade 2</th>
<th>Grade 3</th>
<th>Grade 4</th>
<th>Grade 5</th>
<th>Grade 6</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>N = 273 teachers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Much too difficult</td>
<td>16.1 %</td>
<td>16.7 %</td>
<td>15.8 %</td>
<td>10.1 %</td>
<td>12.4 %</td>
<td>8.5 %</td>
<td>13.3 %</td>
</tr>
<tr>
<td>B. Somewhat too difficult</td>
<td>28.5 %</td>
<td>29.1 %</td>
<td>29.2 %</td>
<td>27.3 %</td>
<td>35.4 %</td>
<td>38.3 %</td>
<td>31.3 %</td>
</tr>
<tr>
<td>C. Well adapted to the grade</td>
<td>55.3 %</td>
<td>53.2 %</td>
<td>52.9 %</td>
<td>60.4 %</td>
<td>50.8 %</td>
<td>48.9 %</td>
<td>53.6 %</td>
</tr>
<tr>
<td>D. Somewhat too easy, or too easy</td>
<td>0.2 %</td>
<td>1.1 %</td>
<td>2.1 %</td>
<td>2.1 %</td>
<td>1.5 %</td>
<td>4.4 %</td>
<td>1.9 %</td>
</tr>
<tr>
<td>TOTAL</td>
<td>100 %</td>
<td>100 %</td>
<td>100 %</td>
<td>100 %</td>
<td>100 %</td>
<td>100 %</td>
<td>100 %</td>
</tr>
</tbody>
</table>
The teachers were virtually unanimous in confirming the absence of “lax” exams and the overly demanding character of many exam questions. The questions that were judged as “too easy” represent a negligible minority (1 to 2% in the lower cycles of primary education, 4 to 5% in the higher cycle), while the questions considered as “too difficult” represent between a third and 45% of the content for mathematics exams and between 40 and 48% for French exams. On average, therefore, only 60% of the Maths questions and scarcely more than half of the French ones were considered by the teachers to be “well adapted” for the grade in which they were used.

The consistency in the proportion of questions considered to be “suitable” or “unsuitable” for each grade, was not the only one. In general, real agreement could also be noted on the type of questions that were considered to be well adapted and, additionally, on the implicit or explicit criteria that the teachers used to accept or reject a particular exam question. The questions considered as suitable were generally so judged by more than 70% of the teachers; and the comments that accompanied their rejection showed, generally, that the reasons were similar (the task was too artificial, it was outside the curriculum, it included too many difficulties combined in the same item, it was strewn with useless “pitfalls”, or it was too long, etc.).

The opinions concerning the length of the exam questions were, for example, impressively precise. Confronted with a series of twenty written calculation, spelling or vocabulary questions, similar in nature, but including unequal number of items, the teachers were unfailingly consistent. Their answers made it possible to identify, in each case (and almost to the nearest unit!) the number of items that a “reasonable” question should include and beyond which one could not any longer require that the question be “entirely or almost entirely solved” by students in a given grade.

When questioned apart from their class context, the teachers demonstrated substantive professional knowledge about achievement standards that can be required. Their precision and consistency contrasted in a surprising way with the practices that they used when developing their own classroom exams.

On the basis of these opinions, and in collaboration with groups of volunteer teachers, the exam tabulations were reworked. For each curriculum objective, only those questions considered to be “well suited” by at least 75% of the consulted teachers were retained. As needed, the instructions or the wording of the items were re-examined. When it appeared that important curriculum objectives were insufficiently represented (in French, for example, this was the case for listening comprehension and written expression), then some new questions of equivalent difficulty were created to allow their assessment. The layout of the questions, their formatting, and the quality of their graphics and illustrations were improved.

A series of reference documents containing, for each grade and each curriculum objective, a choice of questions considered to be “reasonable” by a majority of teachers teaching that grade was thus developed and circulated to all the participants. They were invited to use them to develop their future classroom exams, either by borrowing questions from them or for inspiration when creating questions of equivalent difficulty.

The comments collected following the distribution of these documents were very positive, and showed that the action taken touched a sensitive nerve among the teachers: the anxious concern to find out, as one of them said, “whether we are on the right track”, i.e. whether we are asking enough of the students, whether some requirements are not too limited compared...
to what is done in other classes or other schools. To discover that the “reasonable” assessment standards represented by the questions included in the reference document were, in fact, shared by a very large majority of colleagues provided each one with the impression that it was after all possible to guide all students to an acceptable proficiency level.

Significantly, the impression that dominated in the opinions expressed by the teachers was one of relief (“I was reassured about the level of my work in grade 2”; “What a professional reassurance!”; “A wealth of reassuring information”). The majority of the teachers said that they used these documents as the basis for preparing exams, but often also for changing somehow the focus of their work in class and their assessments throughout the school year.

The reference documents were a success, not only among staff in the schools that participated in the studies, but in many other schools. Several years after the conclusion of the study, they remained one of the most asked for SPE publications.

The evaluation of the intervention done at the end of the study indicated that the retention rates were somewhat reduced in the participating schools. At the beginning of the intervention, 28.7% of students were behind the grade in the total population of the sampled schools. After six years, when the study was completed, the rate was 18.7 %. The reduction trend was observed in the great majority of the involved schools, though its amplitude varied (Table 7).

Table 7. Evolution of the percentage of grade-retained students in the sampled schools

<table>
<thead>
<tr>
<th>Percentage Decrease</th>
<th>Number of Schools</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 to 20 % decrease</td>
<td>10 schools</td>
</tr>
<tr>
<td>10 to 15 % decrease</td>
<td>8 schools</td>
</tr>
<tr>
<td>5 to 10 % decrease</td>
<td>18 schools</td>
</tr>
<tr>
<td>1 to 5 % decrease</td>
<td>9 schools</td>
</tr>
<tr>
<td>1 to 3 % increase</td>
<td>4 schools</td>
</tr>
<tr>
<td>Data not available</td>
<td>3 schools</td>
</tr>
</tbody>
</table>

This positive evolution was not only a result of the intervention. A law limiting the possibility of making students repeat a grade more than twice during their primary education was passed about at the same time, which certainly favoured the trend towards reduction. Indeed, during the same period, a reduction in the percentage of grade-retained students was also observed in schools not involved in the intervention, although to a lesser degree (-5.8% on average, vs -10% in the participating schools).

Such a result appears quite modest, given the duration of the intervention and the relatively significant resources that were invested in it. In particular, the action did not seem to have managed to really destabilise the belief, well anchored among the majority of the teachers, that repeating is a necessary evil and that it is impossible to manage a too heterogeneous class without risking a downward levelling. This “school failure culture” remains powerful. It leads French-speaking Belgian teachers to find natural, all things considered, that about one student in five ‘can not learn’ in six years a curriculum that was devised as the foundation of further schooling for the entire population.
And at the secondary level?

It has been seen above that grade repetition only slightly reduces the heterogeneity faced by the institution, and that the poorly monitored efforts of teachers to adapt the curriculum to their students’ level produce many negative effects. These are supported by the competitive context characterising the educational system in the French Community of Belgium: a school or a class with no repeating would be quickly regarded as “lax”. “Adaptation” thus mainly translates into a generalised upward departure from the curriculum, most evident in the strongest classes, somewhat less in the weakest, but which renders critical, in all of them, the situation of those students who are deprived by such overloads of the time they need to attain a sufficient and lasting mastery of the subjects taught.

In the first year of secondary education, the Belgian system includes “transition classes” for students who completed their primary education without attaining a sufficient proficiency level. The curriculum in such classes is simplified and oriented towards bringing up to level and motivating these students in difficulty. One might think that this differentiated curriculum offered to the most vulnerable students would result, at the very least, in less heterogeneity and less disparity in requirements in “ordinary” classes.

However, the studies conducted at this level (Detheux, 1992, Detheux and Kellens, 1992) appear to establish that this does not happen. The authors observed departures from the curriculum comparable to or even exceeding those observed in primary education. Just as their colleagues in primary schools, the teachers in a sample of secondary schools were invited to comment about exam questions that they themselves developed for the end of the first year of secondary education. They found that only 25% of the question in the French exam papers were at an “appropriate” level of difficulty, with the others being “too difficult”. The proportion was 22% for English as a foreign language. Besides, as in primary education, the marking system tended to overestimate the heterogeneity of each class and to overlook differences between classes. Again, grade repetition rates were extremely high in these schools.

Conclusion

The educational system of the French Community of Belgium undoubtedly constitutes a somewhat extreme case in respect of the issue under consideration. It however makes it possible to better understand some of the mechanisms that make repeating so ineffective as a “remedial” measure:

- It favours the emergence of a concept of school failure that tends, to a certain extent, to disempower teachers, by persuading them that it is “normal” that certain students should fail;

- In the absence of a sufficiently clear official definition of what “success” should be, it encourages teachers to adapt their requirements to the level of their class - or rather, to adapt them to the level of the group of the best students in their class. Unclear standards of success maintain, indeed, the insecurity hinted at by the situation
described above, from which teachers protect themselves by developing exams that are systematically more difficult than required.

- Under these conditions, and even when an official curriculum differentiation is available (we have seen that this is the case in the Belgian secondary schools), each class produces, to some extent, its own failures.

- Through a perverse paradox, success for all is no longer (and can no longer be) the goal to be pursued: it would be interpreted by colleagues and students’ parents as a sign of an unacceptable relaxation of requirements. A high rate of repeating is, on the other hand, by no means disqualifying for a school. The educational community tends even to view it as a positive criterion, indicating that the teaching provided by the most ‘severe’ schools is of a quality without concessions.

Nowadays, it is widely held that the fight against school failure must be addressed through a greater differentiation of teaching and improved adaptation to local needs. In France and in Belgium (two countries where the practice of repeating is maintained), this is translated, in particular, into directives encouraging schools to develop specific educational projects, adapted to their population. The studies presented above may cast doubts on the appropriateness of these policies in school systems where the disparities between individual schools are already very significant, unless strong monitoring is set up to avoid a possible corresponding increase in the disparity of curriculum requirements.

While it is desirable, indeed, to practise in each class a differentiated pedagogy and to avoid frontal teaching that is “indifferent to the differences” (Perrenoud, 1982), it is also true that the disparities between classes and schools can unleash a mechanism that reinjects inter-individual inequality each time students cross one of the multiple partitions of the system. To fight against failure is thus Sisyphean work as long as the two sources of inequality are not taken into account, and each one fought with appropriate means.

To homogenize the objectives to be reached, through curriculum clarification, through a more clear definition of standards of success, and through more attentive assessment of adherence to these, appears, from this perspective, to be at least as urgent as diversifying the paths leading to attaining those objectives.

References


PERRENOUD, P., (1982), L’inégalité quotidienne devant le système d’enseignement. L’action pédagogique et la différence, Revue européenne de sciences sociales, 63, 87-142.


