An Examination of Differences in High School Graduation Standards

Philip Nagy[1]
ontario institute for studies in education

This study describes the method by which Newfoundland identifies discrepancies among high school graduation marks and the consequences for students of adjustments made to reduce these differences. Analysis is based on results of the 1980 high school graduating classes, where a system of joint evaluation allows a comparison to be made between school marks and provincial examinations. The analysis focuses on: (a) the size and distribution of differences; (b) whether they are related to size of school; (c) how much of the phenomenon can be attributed to random error; and (d) whether the discrepancies are primarily a function of schools or individual teachers. A second aspect of the paper considers: (a) the assumptions necessary to perform the analyses herein; (b) the assumptions necessary to link mark discrepancies to teacher generosity; and (c) the fairness of using measures such as these to adjust student grades.

Cette étude touche au problème de la remise des diplômes lorsque les critères d'évaluation sont différents selon les écoles, et des injustices que cette façon d'évaluer peut engendrer chez les finissants. Une étude statistique des résultats de 1980 chez les finissants des écoles secondaires de la province de Terre-Neuve, nous démontre l'ampleur de la situation. Dans cette province, l'évaluation est partagée entre un examen provincial et d'autres examens relevant des écoles secondaires locales. Le rapport fait le point sur: (a) l'étendue du problème; (b) s'il est rattaché au calibre de l'école; (c) dans quelle mesure ce phénomène est rattaché à des erreurs dues au hasard; (d) si ces divergences entre les standards sont tributaires de l'école ou des enseignants en particulier. Enfin, trois éventualités sont examinées: premièrement, l'évaluation de la compétence des professeurs; deuxièmement, la crédibilité de ce genre d'analyse; troisièmement, l'impartialité des vérifications de notes selon les indications données par le professeur.

An ongoing problem in secondary education is the fact that grades awarded by different high schools may reflect different standards and expectations, resulting in difficulty in comparing students for post-secondary admission. This has manifested itself in the establishment of entrance exams for university, in the adoption of system-wide standardized testing programs, and in the current interest in competency testing. As each of these proposed solutions to the problem raises further concerns, care must be taken in choosing remedial action. The purpose of this study is to assess the magnitude of high school graduation grade discrepancies, so that debate may be conducted on the basis of accurate data rather than subjective impressions.

Admission to American universities with competitive admission, and to limited enrolment programs in otherwise free-access universities, is decided, by and large, on some combination of Scholastic Aptitude Test scores, personal interview, recommendation and, perhaps, rank in high school graduating class. The average grade awarded by the high school rarely enters the picture, and probably rightly so. In contrast, Canadian universities, until relatively recently, have based the academic aspects of their admission procedures on high school grades. Equity was assured, in a relative sense, by the use of provincial examinations at the end of high school. However, as these exams have been phased out over the past two decades, concerns have arisen for equity of post-secondary admission procedures and, not incidentally, quality of education. Several universities have considered rating high schools, and the use of placement and/or entrance exams is growing. The data of this present study offer an opportunity to examine the differences (under certain assumptions to be discussed) in expectations or standards across a set of high schools within one province.

In the early 1970s, Newfoundland abandoned the practice of assessing graduation solely by external examination. However, unlike many other provinces, rather than drop the exams entirely, a system of shared evaluation was introduced, with the school responsible for one half the mark, and the other half decided by a provincial exam. Given the availability of two marks for every student, it is possible to calculate an index of the discrepancy between marks awarded to the students of a school by the school itself and by the provincial examination system. This article examines the following issues: (1) How widespread are discrepancies in high school graduation grades? (2) Can the phenomenon be associated with smaller, more isolated schools? (3) How much of the phenomenon can be attributed to random error? (4) Are variations in grades teacher-related or school-related?

We begin with a description of the index of discrepancy, followed by a discussion of assumptions concerning the distribution of student grades needed for the analysis to proceed. Specifically, there are two questions at issue: whether indices of mark discrepancy can be pooled across school subjects; and how the estimation of random error necessary for question 3 above should be carried out.

In the discussion two major issues are addressed. First what are the assumptions under which the mark discrepancies may be viewed as measures of teacher generosity with grades? Second, how fair is the use of such measures for the adjustment of student grades to a common standard? During the reporting, however, it will be convenient to talk of mark discrepancies as differences in "generosity," with quotation marks implied.

The order of procedure, then, will be to address the questions in the order given. Between questions 2 and 3 will be inserted data pertaining to pooling of indices across subjects.

THE DISCREPANCY MEASURE

Since the introduction of the shared evaluation system, the Newfoundland Department of Education has established an index "D," as a check on the consistency with which schools award their share of the final high school mark. (When these data were obtained, Newfoundland high school ended after Grade 11. Since that time, the program has been extended by one year.) Under the assumption that students within a school should, on average, do as well with respect to the province on school evaluation as on the provincial exams, the value D expresses, for each subject, the relative generosity of each school with awarded marks.

The value D is calculated, for each school-subject combination, as:

$$D = (SS - PS) - (SP - PP), \tag{1}$$

where SS is the school average on the school evaluation, PS is the provincial average on all school evaluations, SP is the school average on the provincial exam, and PP is the provincial average on the provincial exam. D is expressed in terms of the differences between the two school averages and the corresponding provincial averages. Thus, the magnitude of D is unaffected by differences at the provincial level between school assessment averages and exam averages. If the expression for D is rearranged to give

$$D = (SS - SP) - (PS - PP), \tag{2}$$

then it is readily seen that the expression in the right-hand set of brackets is a constant for all schools in a given subject, and D is the school deviation about that provincial constant. Also, the expression in the right-hand brackets is the weighted (by enrolment) average of those in the left-hand bracket over all schools. Less generous schools are characterized by negative D values, and more generous by positive ones.

For example, if, for a given school and subject, SS = 66, PS = 60, SP = 70 and PP = 58, equation (1) or (2) yields D = −6. That is, students were assessed as being six marks above the provincial average of school marks by their own school assessment system, but came out as being twelve marks above the provincial average on the provincial exam; the school has been six marks under the average in generosity. Such a judgment assumes that the difference between the school and provincial averages on the provincial exam (SP − PP) is an accurate measure of the school's "true" placement with respect to a provincial standard, and therefore that a school which places its students in a different position with respect to the provincial average of school marks (SS − PS) is, averaged over all students in the class, assessing inaccurately. Thus, in the above example, even though the school awarded marks above the provincial average of other schools, it is judged below the average in generosity because, on the provincial exam, its students were even further above the provincial average.

A decision was taken by Newfoundland authorities to modify school marks by the subtraction of D from all school marks in a school. Thus, marks from more stringent schools would be raised, and those from less stringent lowered. Since only the school marks are adjusted, and not the exam marks, the overall effect would be to adjust the students' final marks by one-half the D value, in the appropriate direction. In the years 1979–1981, officials in Newfoundland adjusted school marks only to within a tolerance of ± five points.

The justice of such a modification will be discussed below, but for now a word of caution about the results is in order. Student grades are reported as submitted by the schools, as if they were unchanged by provincial officials. Thus, we are able to report the magnitude of mark differences before becoming involved in a debate on the meaning of "generosity" and on the fairness of provincial interference in the school marking process. As a consequence of ignoring, in a sense, that marks were actually changed, however, we report what would have happened without intervention as if it had actually happened. The occurrence of intervention in no way interferes with the accuracy of our estimates of discrepancies, except in one way. A major limitation of these data is that they were produced by teachers who knew in advance that their submitted marks would be, in fact, subject to modification. Thus, to the extent that mark discrepancies are deliberate, rather than the result of lack of experience or random error, our data are affected by changes in the motivations of teachers caused by imposition of the modification scheme.

RANDOM ERROR AND POOLING

Much of the analysis reported here depends on estimating the variation in D to be expected by chance. This, in turn, depends on the viability of pooling D across subjects. In many cases, within a particular minority interest subject, there are not enough classes to put much faith in the distribution of D within that one subject. Thus, pooling is put forward as necessary for stability of estimates. Also, the variability in D due to chance alone is a function of the class size group under discussion, and with wide fluctuation in the popularity of subjects, class size is not uniformly distributed across subject choices. Thus, except for the large enrolment subjects, there is little option but to pool Ds. In order to proceed down this path, we begin by creating an individual analogue of D, D_I, for computational convenience. Thus:

$$D_I = (IS - PS) - (IP - PP), \tag{3}$$

where IS and IP are the individual marks from the school and the provincial exam respectively, and PS and PP are as above. Rearranging gives:

$$D_I = (IS - IP) - (PS - PP), \tag{4}$$

where, as in equation (2), the right hand bracketed expression is a constant for the province.

Summing across N individuals within a school-subject combination produces:

$$\Sigma D_I = \Sigma IS - \Sigma IP - N(PS - PP), \tag{5}$$

and averaging gives:

$$\frac{\Sigma D_I}{N} = \frac{\Sigma IS}{N} - \frac{\Sigma IP}{N} - (PS - PP)$$

But,

$$\frac{\Sigma IS}{N} = SS \text{ and } \frac{\Sigma IP}{N} = SP$$

and therefore, from equation (2),

$$D = \left(\frac{1}{N}\right) \Sigma D_I \tag{6}$$

In order to assess the effects of random error, it is necessary to conceive of groups of students taking a particular subject in a particular school as nothing more than a random sample of mark discrepancies D_I, independent of any systematic teacher or subject choice effects. This conceptualizaton does not reflect reality, of course. Students are in particular classes due to subject choice and location of residence, and their D_Is are not independent, in that they have been assigned by the same teacher (or, at least, the same school, an issue which will be discussed). However, the question is, in effect, if we dismiss all the systematic reasons why D is more than the average of D_Is chosen randomly from a large pool, how large an effect for D can we expect for random reasons alone?

The issue is not how much D_I would vary by chance, but D, the class average of the D_Is. Since (PS − PP) is a constant for the province, we have from equation (4):

$$\sigma_{D_I}^2 = \sigma^2_{(IS-IP)} \tag{7}$$

which yields a convenient mechanism for calculation.

The variance of D_I, at the individual level, is directly related to the variance of D at the class level, by:

$$\sigma_D^2 = \sigma_{D_I}^2/N \tag{8}$$

if the class grouping on which D is calculated represents nothing more than a random selection of size N from the pool of scores. That is, a standard error of the mean for D can be calculated, and an estimate made of the amount of variation in D to be expected by chance, for classes of a given size. This can be compared to the distribution of D for given sizes of classes, pooled across subjects.

It is important to note that the use of D_I does not mean that the distribution of individual marks is being used in place of the distribution of group averages in order to assess the size of random error effects. Rather, the usefulness of D_I is that, through equations (7) and (8), it can be used as a computational convenience in estimating random effects on D. The final comparison is still the expected distribution of class D values, considered as the group mean D_I, under artificial random assumptions compared to actual distributions under real conditions.

The above discussion assumes that pooling across subjects is reasonable. This issue is open to empirical examination. If we are to imagine drawing samples of various sizes from a large pool of individual D_Is, then to justify ignoring subject differences, we must show that the D_Is are distributed in like manner for different subjects.

Due to the large sizes of many subject enrolments, any statistical test applied to the variances of individual subjects compared to each other or to a pooled value will turn up many significant differences, suggesting that the pooling procedure should be questioned. However, under the circumstances, the reasonableness of pooling may be examined by comparing the subject enrolment in a "slice" of the pooled distribution with that in the entire distribution.

The procedure used is as follows. First, the pooled standard deviation of D_I is used, with the normal table, to demarcate the middle 95% of the pooled distribution. Thus we have cut-off points which exclude the extreme 5% of the pooled distribution. Using the standard deviations of D_I for individual subjects and these same cut-off points from the pooled distribution, we can calculate what proportion of the D_Is for each individual subject are outside the cut-off points based on the pooled distribution. For those with smaller standard deviations, this will be less than 5%; for those with larger, more than 5%. In order to graphically reflect different subject enrolments, these proportions are combined with subject enrolments, and the reported comparison is the percentage of the total pool for each subject with the percentage of the extreme 5% of the pool for each subject.

The purpose of pooling, then, is to allow use of equation (8) and the t-distribution to assess how much variation in D can be expected by chance alone for different size classes. Working from the proportion of classes within these bounds and the class sizes, we can estimate the proportion of students whose grades have been affected by more than could be expected based on chance alone. There is no suggestion that chance is ever the only factor involved, or that we have identified individual cases. All we have is a comparison of the size of the actual discrepancy phenomenon with that predicted on the basis of assumptions of random error alone.

SAMPLE AND METHOD OF ANALYSIS

Analysis began with all unmodified Grade 11 marks obtained in the Newfoundland provincial exams in June 1980 by students in schools on the shared evaluation plan. (A few isolated schools are not on shared evaluation; one such school, with only nine students, was mistakenly included in the analysis.) An attempt was made to remove from the file all part-time students, considered as those who did not take English, the only compulsory subject for all students. Some part-time students who did take English undoubtedly remain in the sample, but their numbers are small. Provincial and school averages, and Ds, all rounded to the nearest whole number, were calculated for the 23 subjects for which provincial examinations were offered.

One of the objections raised by teachers to the mark modification scheme is that, in small enrolment subjects and small schools, one or two isolated individuals can have an inordinate effect on the D factor used to modify their classmates' grades. In the interests of fairness, the Newfoundland officials introduced a scheme to identify such aberrant individuals, and remove them from the calculations. Although the exact method has varied a bit, essentially, the method of removal involves examination of the personal discrepancy factor of each student within each subject, and removal of anyone with a factor over a large positive or negative number. In our analysis, this was not done. Fewer than 200 of the 42 265 grades fell into this category. Given that the purpose of this study was to document the size of discrepancies and not to improve the fairness of mark adjustments, it was not considered to be worth the computing costs for a relatively minor improvement in the accuracy of the provincial picture. Hand analysis of some of these situations showed that: (a) most of the really extreme D factors were not affected by this adjustment; (b) removal of the aberrant students usually did not change the D by more than one or two. Thus, the most obvious danger of our analysis shortcut, that we are reporting on many extreme mark modifications which were not, in fact, done, is not a concern. While there are undoubtedly a handful of results among our more than 42 000 marks which do not correspond to the actual student marks, they do not affect the provincial picture either significantly or in any particular direction.

After calculation of D, student marks were recalculated on the basis of removal of differences in school marks due to D. That is, marks were set to what they would be if, in effect, D were zero for all schools offering a subject. Returning to the above example, with a D of −6, if an individual student had been given 74 by the school and earned 68 on the provincial exam, his unadjusted final mark would be 71. In the adjustment, his school mark would be raised to 80, his exam mark would remain the same, and his adjusted final mark would be 74.

Reporting is from the perspective that this second set of marks is "correct," while the original marks are in "error." These two terms are, of course, only a convenience. There were 63 cases of marks pushed over 100 by recalculation. These were all changed to 99. Final marks of 48 and 49 were changed to 50.

Next, students' status and averages were calculated for each set of marks on the basis of three criteria: first, pass/fail, 50% on English and any four other credits; second, eligibility for university, 60% on English and five other credits, subject to certain subject-choice restrictions which need not concern us here; and third, honours, 75% on the same subject-choice criteria as admission to university. For ease of computation, we did not drop the lowest subject for students who had taken more than the minimum number of credits. This is a small source of error in our data.

RESULTS

Magnitude and school size

The data in Tables 1 and 2 set the context for the results. Table 1 contains the class sizes of the 174 schools in the analysis, expressed in terms of the size of the group writing the English exam. As can be seen, the majority of the schools in Newfoundland are quite small by urban Canadian standards. Only a handful have more than 100 students in their graduating classes. Such a sparse population does not help in generalizing these results to more densely populated parts of the country, but if we can hypothesize that teacher isolation, with accompanying lack of perspective on a provincial standard, is part of the cause of the phenomenon under investigation, then these data should help to illuminate the issues. Table 2 contains subject enrolments.

Table 3 reports the extent of the discrepancy in school-awarded grades, and addresses the first of the major questions outlined above. If we consider (somewhat arbitrarily) that variations in generosity of more than five marks in either direction are serious, then about 39% of student grades fall into this category. Comparing the bottom two and top two rows of the table, there seems to be a slight trend for more students to have been treated extremely generously compared to those treated extremely ungenerously: there were 5.4% of students with marks inflated by more than ten, and only 4.0% with marks deflated by a corresponding amount. A second trend noticeable from these data is that classes with extreme D values tend, on the average, to be small. There are 13.8% of classes in the four extreme categories, but these classes contain only 9.4% of the students.

TABLE 1
Summary of School Sizes
(by English Enrolment)

School Size (Students)	*Number of Schools*	*Percentage of Schools*	*Percentage of Pupils*
2– 10	26	14.9	2.1
11– 40	45	25.9	9.2
21– 40	43	24.7	16.1
41– 60	27	15.5	17.4
61–100	17	9.8	15.7
101–200	11	6.3	20.8
>200	5	2.9	18.9

TABLE 2
Summary of Subject Enrolments

Subject	*Number of Schools*	*Number of Students*	*Subject*	*Number of Schools*	*Number of Students*
English[1]	174	7637	Economics[1]	150	4845
Honours Math	31	651	R.C. Religion[2]	64	2843
Matric Math[1]	173	5503	World Religion[2]	29	594
Basic Math[1]	110	1117	Integrated Religion	9	134
Physics	82	1606	Pentecostal Religion	9	191
Chemistry	47	938	French[1]	116	2151
Earth Science[1]	62	1231	Art	20	253
Physical Science	36	598	Music	19	148
Biology	135	4121	Home Economics	34	607
History[1]	130	2953	Industrial Arts[2]	8	104
World Problems	14	322	Physical Education	20	429
Geography[1]	118	3289	Total		42265

[1] Includes one school ($N = 9$) of full provincial exam evaluation.
[2] Some missing exam marks (up to 4).

These trends, tentative as they are, suggest some related hypotheses concerning the origins of discrepancies in marks. These suggestions are put forward now, and will be explored below in light of further analysis. First, the trend for smaller classes to have more extreme D values in both directions suggests that teachers in smaller schools have more variation in their expectations than teachers in larger schools. This may be due to smaller classes offering teachers less opportunity to experience a full range of student ability levels, resulting in less accurate perceptions of what constitutes a provincial standard of work. Also, there may be a link with the trend for smaller communities, which contain smaller schools, to

TABLE 3
Distribution of D (pooled across all subjects) (%)

Size of Discrepancy	*Classes[2] Affected*		*Individual Grades Affected[3]*	
Less than −15	(22)	1.4	(410)	1.0
−11 to −15	(62)	4.0	(1240)	3.0
−6 to −10	(241)	15.6	(6800)	16.3
−3 to −5	(213)	13.8	(6440)	15.4
−2 to +2	(454)	29.4	(12845)	30.7
+3 to +5	(226)	14.6	(6355)	15.2
+6 to +10	(198)	12.8	(5420)	13.0
+11 to +15	(105)	6.6	(1880)	4.5
More than +15	(25)	1.6	(390)	0.9
Totals	1546		41780	

[1] Negative values indicate less generosity than average; positive more generosity.
[2] A 'class' refers to all students in a school taking a subject. In large schools there would be several sections to a class, and often more than one teacher.
[3] Estimates based on the assumption that all classes in a size range of 10 are at the midpoint of that range. For example, four classes in the size range 21–30 contain 4 × 25.5 = 102 people. Total does not coincide with Table 2.
NOTE: *N* is in parentheses.

attract, on average, less experienced teachers. Second, there is a traditional tendency for school achievement in Newfoundland to be lower in less urban, more isolated communities, and therefore in smaller schools (Nagy, Drost and Banfield, 1983). If we can picture two competing tendencies in teachers, to teach and grade the students in their class, and to teach and grade to a perception of a provincial standard, then we would expect that classes would tend to be graded as closer to the average than they really are: that is, weaker classes would be given higher marks than they deserve on a provincial standard, and better classes would be treated in the opposite manner. It would also be consistent to suggest that teachers of smaller classes are more able to gear teaching to individuals. This might result in more individualized and better teaching, but could also cause deviations from a provincial perspective in grading.

As the discrepancy factor is defined, the provincial average D, over all students taking a subject, would be zero, except for rounding error. These hypotheses, therefore, taken together, predict a pattern such as that found: smaller schools with most of the extreme D values, on balance with a slightly positive D, as weaker students are dealt with generously, and larger schools with less extreme D values, on balance slightly negative, as better students are treated less generously. There are, of course, many exceptions to these general trends, as large schools contain many poor students, and small many good ones.

Table 4 reports the breakdown of the discrepancies by class size. For

TABLE 4

Distribution of D by Class[1] Size – Percentages of Total Individual Marks

Class Size (number of students)	*Less than −15*	*−15 11 to*	*−10 to −6*	*−5 to −3*	*−2 to +2*	*+3 to +5*	*+6 to +10*	*+11 to+15*	*More than +15*	*Total*	*Average D*
1– 10	0.13	0.29	0.84	0.61	1.34	0.62	0.71	0.50	0.15	5.19	+0.40
11– 20	0.18	0.58	2.85	2.52	5.11	2.85	2.33	1.31	0.22	17.95	+0.44
21– 40	0.27	1.03	3.83	3.73	8.41	3.75	2.73	1.40	0.44	25.59	0.00
41– 60	0	0.95	1.68	2.30	5.25	2.06	2.20	0.49	0.11	15.04	−0.06
61–100	0.39	0.36	2.58	1.50	4.66	3.27	1.92	0.34	0	15.02	−0.36
101–200	0	0	1.92	2.77	4.11	1.52	2.30	0.40	0	13.02	0.25
200+	0	0	2.54	1.98	1.81	1.12	0.74	0	0	8.19	−2.18
Total	0.97	3.21	16.24	15.41	30.69	15.19	12.93	4.4	0.92	100.0	−0.11

[1] Class refers to all students in a school taking a subject.

example, 0.13% of the student grades awarded were to students in classes of ten or fewer students with a D value of less than −15. Some of these students are from classes which might be low enrolment subjects in fairly large schools. Note that, to arrive at these figures, two approximations were used, based on actual counts of classes rather than students: one, as explained in the footnote to Table 3, that classes in a size range were at its midpoint; and two, that classes in a D-range were at its midpoint. For improved accuracy, when a size range reported is greater than ten, separate estimates were made for each size range of ten, and then summed. All estimates have been rounded to the nearest five.

Table 4 shows the association between class size and discrepancy from the provincial average in generosity. As class size increases, fewer students are subjected to extreme deviations from the average. Fewer than 10% of grades have a D greater than 10 in either direction and the majority of these occur in small classes. The column totals of Table 4 show that about 60% of marks have a D value within ± 5. While these bounds have no statistical basis, they do have an intuitive appeal as reasonable tolerances. In fact, this range is the allowable deviation decided upon by the Newfoundland Department of Education.

Before proceeding, a comparison of the right hand column of Table 1 with the second column from the right in Table 4 shows that average class size is considerably smaller over all subjects, optional and compulsory, than in the compulsory English classes. While 8% of students take English in groups of 20 or fewer, 23% of students take all subjects in classes of this size. This makes the issue of class size and its relationship to random error in the grading process quite important to a fuller understanding of these results. We shall return to this issue after an aside.

The aside focuses on the right hand column of Table 4, which gives the average D for classes of a given size. This confirms the trend suggested in the discussion of Table 3, that small classes tend to have slightly positive D values, and larger slightly negative. The bottom right hand entry in the table, −0.11, is the overall average D value for all grades awarded, and should theoretically be zero. The fact that it is not is due mostly to error introduced by our approximation methods and partly to the rounding of marks in calculating school averages, provincial averages, and D, and to the raising of marks of 48 and 49 to 50.

Random error and pooling

Before addressing question 3, on the random error of D, we must first present data on the viability of pooling D values across subjects. Table 5 presents data which will allow consideration of this issue. The second column of this table contains the standard deviation of D_I for each subject as calculated through equation (7). These values may be compared with the pooled value of 11.2. Note that the distribution of D, as might be

TABLE 5
Data on Pooling Across Subjects

Subject	Standard Dev. of D_I	Percent of Total Distribution of D_I	Percent of 5% Fringe of Dist. of D_I	Correlation (IS and IP)
English	9.6	18.1	8.2	0.68
Honours Math	11.1	1.5	1.5	0.75
Matric Math	11.4	13.0	13.8	0.82
Basic Math	12.0	2.6	3.5	0.71
Physics	11.5	3.8	4.2	0.76
Chemistry	11.5	2.2	2.5	0.80
Earth Science	12.0	2.9	3.9	0.73
Physical Science	12.5	1.4	2.2	0.66
Biology	10.3	9.8	6.4	0.81
History	14.2	7.0	16.8	0.76
World Problems	12.3	0.8	1.1	0.78
Geography	10.7	7.8	6.2	0.75
Economics	10.7	11.5	9.2	0.75
R.C. Religion	12.5	6.7	10.4	0.72
World Religion	11.0	1.4	1.3	0.74
Integrated Religion	10.6	0.3	0.2	0.81
Pentecostal Religion	10.9	0.5	0.4	0.81
French	10.2	5.1	3.2	0.76
Art	13.2	0.6	1.2	0.63
Music	15.8	0.4	1.1	0.69
Home Economics	10.8	1.4	1.2	–[1]
Industrial Arts	9.6	0.2	0.1	–[1]
Physical Education	12.0	1.0	1.4	–[1]

[1] Not available due to data storage methods.

implied from Table 3, is, as expected, narrower than that of D_I here. The next two columns of Table 5 may be easily misinterpreted. An example clarifies their calculation.

The 95% confidence interval of 1.96 standard deviations becomes, for the pooled distribution, 21.9 (1.96 × 11.2) marks. This number of marks corresponds to 2.28 (21.9/9.6) standard deviations of the English mark distribution. Thus, while 5% of all marks can be expected to have D of more than 21.9 marks, by chance alone only 2.26% of the English marks can be expected to deviate that much due to chance alone. That is, 2.26% of the normal distribution is more than 2.28 standard deviations from the mean. Thus, 173 (.0226 × 7673) English marks can be expected to have an associated D greater than 21.9 marks by chance alone. In the pooled distribution, we expect 5% of 42 265, 2 113 marks in this category. Thus, the English marks represent only 8.2% (173/2113) of the marks with D_I

greater than chance, although, in total they represent 18.1% of all the marks.

These two columns, then, give, for each subject, the percentages of the D_Is in the total distribution and in the 5% fringe. If the distribution of D_I is independent of subject, then the two columns should be the same. That is, any slice of the distribution should contain the same proportion of each subject's D_Is as any other. To the extent that the column values match, the pooling of D_I (and D) across subjects is justified. In fact, the case chosen for explanation, English, is the worst match of all subjects. In 16 of the 23, match is within one percent. Note that, since the purpose of the comparison is pooling, the proper comparison is subtraction, rather than division. English, then, is quite underrepresented in the tails of the distribution, while, at the other extreme, history is quite overrepresented.

One is free to decide whether the data in these two columns constitute sufficient match to allow pooling of the D_I across subjects. If pooling is rejected, then the implication is not that the conclusions based on pooling are invalid overall, but merely that they might not apply equally well to those subjects which are not proportionately represented in the two columns. Thus, more attention might have to be paid to the fact that some subjects, English and history for example, are graded with different degrees of variation in generosity. That is, the conclusions will hold for each subject inasmuch as the standard deviation of D_I for that subject is close to the pooled value of 11.2.

There is no intent that this portrayal of the data be taken as a legitimate statistical test, but merely as a defence of the reasonableness of the procedure. If a slice had been chosen closer to the middle of the distribution, a better picture could have been painted, and farther out, a poorer one.

The last column of Table 5 presents the correlation between individual school and exam marks for each subject. It is included here for convenience, and will be discussed later.

Using the value of σ_{D_I} for pooled subjects, it is possible to put cut-off points around the expected limit of chance distribution of D_I. For example, using a class size average of 5 for all classes between 2 and 10, we get a standard error of the mean of $11.2/\sqrt{5} = 5.0$. Reference to the t-distribution (df = 4) gives a 95% confidence interval of ± 13.9 marks around the expectation of D_I due to chance, for a group of this size. Under the assumption that class groupings are random selections of a given size from the D_I distribution, 95% of such groups should have D within the range ± 13.9. Thus, we can estimate the proportion of cases in which factors larger than chance are operating at the class level. Noting that mean D_I within a class is simply D, and that the groups are about equal size, we can estimate the proportion of individual marks in which D (not D_I) differs from zero by more than chance. There is, of course, no implication that the individual cases have been identified. Table 6

TABLE 6
Influence of Chance on D

Class Size	95% confidence limits (in marks)	Percentage of Grades: D low by more than chance	D different from zero by less than chance	D high by more than chance
2– 10	13.9	4.3	89.7	5.9
		0.2	*4.6*	*0.3*
11– 20	6.2	17.9	62.4	19.7
		3.2	*11.2*	*3.5*
21– 40	4.2	25.8	50.7	23.5
		6.6	*13.0*	*6.0*
41– 60	3.2	29.0	42.1	28.9
		4.4	*6.3*	*4.3*
61–100	2.5	32.2	30.7	37.1
		4.8	*4.6*	*5.6*
101–200	1.8	40.3	24.4	35.4
		5.2	*3.2*	*4.6*
200+	1.3	60.0	12.6	27.5
		4.9	*1.0*	*2.3*
Percent		*29.4*	*44.0*	*26.6*

NOTE: Numbers in roman indicate percentage within class size range; numbers in italic indicate percentage over all class sizes.

reports the results of these calculations with number of grades awarded expressed first as a percentage within each class size, and second as a percentage of the total group. Within each size group, 5% of cases should lie outside of the chance limits. For every size, the results show substantially more than 5%, indicating that factors larger than chance are at play in the distribution of D among classes. The confidence widths suggest that more tolerance should be given to smaller class Ds, as is expected. That is, a class D based on a small group can be expected to vary more simply due to chance than one based on a larger group of students. Note that these calculations were based initially on class size intervals of only ten students. Except for the first two rows of Table 6, these initial results were summed for ease of presentation.

The initial impression created by these data suggests that small schools are somewhat better at estimating the grades of students on a provincial basis. By contrast, Table 4 suggests that these schools are much worse. Neither of these statements should be taken as direct reflections of the abilities of the teachers in question, either as teachers or graders. To a large extent, Table 4 reflects the simple truth that averages based on small groups tend to vary more than those based on large groups. Conversely,

Table 6 suggests that smaller classes are better at meeting a given standard of accuracy, but we must remember that this standard, in the sense of the size of the confidence intervals, is less stringent. Drawing any stronger implications from the results of this analysis would be putting more faith in the ability of the data to meet the assumptions necessary to the calculations than is warranted. However, while Table 6 is based on statistical inference, Table 4 is not. Within the limits of the approximations already discussed, we have an accurate picture of the sizes of the deviations caused by the grading of several hundred teachers, and we have a rough idea of how much variation is probably unavoidable. This does not mean, however, that those cases where the actual D is within the confidence interval represent good evaluation, with D caused only by chance, and those outside represent poor evaluation, with D caused by some systematic, school- or teacher-dependent factor. Further, it must be realized that in every situation, including D = 0, we have some combination of chance and systematic factors operating, perhaps in concert, perhaps in opposition. However, it can be stated that about 45% of the D variations, at a provincial level, are no larger than might be expected by chance, while 55% are. Finally, it is important to realize that the above discussion assumes that D measures teacher inaccuracy, a point which has not yet been addressed.

School or teacher effect?

While the above results are interesting, and perhaps important from the point of view of individual subjects, discrepancies among teachers in generosity with marks may have little effect on students if D is distributed randomly across subjects within schools. That is, a student might be penalized six marks in English, but get a "bonus" of five in French. The change in his total mark may thus be negligible, and in his average mark, even less. Also, there is little effect on a student whose average grade is changed from 64 to 68, since either will probably get him/her into university, and neither will win awards. These issues have been investigated by calculating average Ds across subjects within schools, weighted by enrolment, and important transitions (e.g., pass/fail, eligible/not eligible for university, and pass/honours) directly caused by variations in generosity. Note that the average D in a school represents twice the change in the grade average of a typical student, as D measures the discrepancy in school-assigned marks only, which are then averaged with the exam marks.

Table 7 contains estimates of the changes in average mark due to mark adjustment, broken down by school size. These are estimates in that, first, average D calculations are based on overall school subject enrolments and do not reflect individual timetables, and two, because we have again assumed that all schools in a size range are at the midpoint of that size

TABLE 7
Magniture of Changes in Individual Averages Due to Adjustments
Estimated Number of Students by School Size

School Size (No. in English Class)	Size of Adjustment[1] (in Average Mark) More than +5	+3 to +5	+1 to +3	−1 to +1	−3 to −1	−5 to −3	more than −5
2– 10	0	15	30	30	5	30	20
11– 20	0	15	210	270	120	60	0
21– 40	0	0	315	340	295	75	0
41– 60	0	0	300	690	355	0	0
61–100	0	0	365	355	395	150	0
101–200	105	0	280	820	115	165	0
>200	0	0	690	765	0	0	0
Total	105	30	2190	3270	1285	480	20
Percentage	1.4	0.4	29.7	44.3	17.4	6.5	0.3

[1] Positive adjustment equals negative one-half average D.

range. Average marks change by one half the average D, but in the opposite direction. From Table 7, fewer than 9% of students have their averages changed by more than three marks in either direction.

It is interesting to examine, by subject, the number of students who have been awarded a passing rather than failing grade due to teacher generosity higher than the provincial average, and vice versa. Over all subjects, teacher generosity awarded 1 210 credits, and lack of generosity withheld 566 credits (by actual count, not estimate). This statement assumes, of course, the validity of assessing teacher generosity using D, and can be interpreted only in terms of the size of the effect at a provincial level, and not as a measure of the justice done in these 1 766 individual situations.

Table 7 indicates that D is not distributed randomly across subjects within schools, for this would imply most school average Ds would be close to zero. Neither does it suggest high correlation in generosity across teachers within schools, for in such a case, the distribution of Table 7 would resemble that of Table 4. Rather, the truth lies between. There seems to be some trend, but not an overwhelming one, for teachers within a school to agree on a level of generosity with marks. This is in keeping with a general hypothesis which holds that D is partly due to policy, and partly to a combination of inexperience and random error.

Table 8 reports the number of students who have been significantly affected (or rather would have been, had the Newfoundland Department of Education not intervened in the grading process) by differences in

TABLE 8
Transitions Due to D (%)

Transition	*Total Students*
Pass → Fail	1.7 (131)
Fail → Pass	4.0 (306)
Pass → Honours	3.3 (252)
Honours → Pass	1.3 (97)
Ineligible → Eligible for University	2.0 (156)
Eligible → Ineligible for University	2.5 (192)

NOTE: *N* is in parentheses.

teacher generosity as measured by D. Approximately 300 students would have passed solely on the basis of teacher generosity above the provincial average, while about 130 would have failed for the opposite reason. While the results of Table 7 might suggest that the problem of lack of uniformity in generosity is a small one, for 430 students, it clearly is a major issue.

DISCUSSION

Two assumptions need to be discussed. The first involves the use of D as a measure of differences in generosity, where the D values are used to inform teachers how they compare to others. The second involves the use of D to adjust teachers' grades to bring them into line with provincial standards.

The rationale for the trend away from central examination systems has been to lessen central curriculum control, and to allow for the assessment of objectives which are not amenable to written examinations (e.g., laboratory skills in sciences.) Thus, for example, if we consider an individual student who excels at lab work, this should be reflected in his assessment relative to his classmates on a mixed or entirely school-based system, but not on a central exam (only) system.

It would be counterproductive to insist, in a mixed system, that D_I should be zero for all students, for then school marks would merely repeat examination marks. However, an argument can be made that D, at the school-subject level should be zero. This argument is that, even while individuals may have relative strengths and weaknesses for different objectives, at the group level, such individual differences should "average out," and the group as a whole should do no better or worse on the objectives assessed by the school grading system than on those assessed by the central exam. For this to be the case, one requirement is that different teachers teach the same proportion of exam-type objectives on their respective halves of the evaluation. For example, if one teacher chooses to spend his/her "half of the time" on objectives which are mostly assessed by the central exam, and another chooses to focus on different but legitimate objectives, then the students of the first teacher will get, in effect, double credit for the time spent on the exam objectives. While the students of the second teacher have received a broader education, which was part of the rationale for the move from central exams, they could not be expected to perform as well on an exam objective as students who have spent more concentrated effort on a narrower set of objectives. This issue, not addressed in this study, is empirically testable, but the methodology would require not only examination of actual teacher practice in deciding on the objectives for their courses, but also examination of the limits of flexibility allowed by Newfoundland Departmental policy in teachers' curricular choices.

A necessary, but not sufficient, requirement for the justification of D as a generosity measure is that the two sets of objectives, school and exam, be highly correlated. This requirement is easily met, as has been demonstrated by Bull et al. (1977). Using data from the 1975 school year, Bull showed that the median within-school correlations between school marks and provincial exam marks were as high as could be expected given the presumed reliabilities of each measure (e.g., 0.61 for English, 0.85 for Algebra, and 0.78 for History). The final column of Table 5 contains corresponding information for this data set, and, as can be seen, the requirement is still met. However, as much as the correlation coefficients from Table 5 suggest that teacher chosen objectives correlate quite well with exam objectives, the issue here is one of concentration of effort and the identity of the two types of objectives,not their correlation.

Another assumption behind the use of D as a measure of generosity is that, relative to the quality of their own work on the examinaton objectives, teachers have done as good a job in teaching the objectives assessable in the school half of the evaluation. For example, if two teachers do an average job on the examination objectives, and one of them does a superb job on the school objectives (chosen, say, due to an unusually strong interest) while the other continues to do average work, then we no longer have a situation in which D is a measure of generosity alone. This argument is not that if students are exposed to teachers of different quality, they should not be compared. It is, rather, that in one case, teaching quality has changed, between the exam and school objectives, while in the other it hasn't, in a manner not reflected in the basis of comparison, the examination objectives. This issue, while empirically testable, albeit with difficulty, is not addressed in this study. This discussion touches on but does not exhaust, the possible reasons for D being something other than a measure of generosity.

The assumptions discussed above, if shown to be valid, might be enough to take the value of D as a measure of overall generosity of schools within each subject, but might not be enough to justify changing the mark of every student in the school by the value D. It is possible to accept the measure D as a measure of deviation from average generosity at the school level, and determine the outcome of variations in generosity in terms of the number of students affected, without enquiring as to whether the particular individuals in question have been treated fairly. In a sense, this is the stand taken in much of this report.

However, apart from issues not addressable by the data of this study, there are some statistical arguments which might suggest that the adjustment of marks is not fair. For example, consider that no allowance was made, in the data reported, for the fact that, in larger schools, there would be more than one teacher of a subject, and that they might differ in generosity. (At the time the Department of Education position was that this was a problem to be solved internally by the school.) Since data reported earlier suggest that the phenomenon is only partly school based, and to some extent due to the individual teacher, use of one D figure for the students of more than one teacher in the same school could lead to difficulties. Unfortunately, the data set does not allow identification of students of different teachers within the same school, but some reasonable estimates can be made.

Using the data in Table 4, assume that if there are likely only two sections of a course in a school, they are taught by the same teacher. If we further assume that the cut-off size for starting a third section of a course is close to 60 students, then Table 4 suggests that some 35% or more of students take subjects when that subject is offered by more than one teacher in the school. This is a difficulty with the use of D in the larger schools, only partly alleviated by the next point, that the value of D can only be at best an estimate, not a measure of generosity.

Students in large schools whose timetables consist entirely of popular subjects will enjoy the benefit of reasonably stable estimates of D, and therefore their teachers' generosity with marks. However, both those who take minority interest subjects and those in small schools do not share this advantage. These students, then, would seem more vulnerable to anomalies caused by the imposition of large sample statistics on a small sample situation, despite attempts to spot anomalies and the use of an appeal procedure. Luckily, most quite small schools are able to offer only the popular subjects, so that few students are in the unfortunate position of taking (provincially) minority interest subjects in smaller schools. Thus, even apart from the curriculum arguments presented above, only those students taking popular subjects in fairly large schools would seem to be relatively protected from statistical violence.

The final concern has to do with the influence on teachers' morale and self-image by an interference in their professional judgment. While undoubtedly there are some teachers either unable or unwilling to assess students fairly and accurately, it seems that, because of these few, most teachers are put in the difficult position of either having to guess how their students will do on the provincial exam, penalize students by being overly stringent with marks, or have their own credibility questioned because of marks lowered by provincial authorities. The data in Tables 7 and 8 suggest that the problem, while noticeable and serious in individual situations, and certainly the cause of some injustice, is not widespread enough to require a provincial rather than an individual solution. The most unattractive consequence of a blanket method is that, by forcing teachers to stay quite close to the objectives assessed by the provincial exam, the initial purpose of abandoning centralized examination systems is defeated.

SUMMARY

This report began with four main questions: how widespread are mark discrepancies; how strong is the school size effect; how much can be attributed to error; and what is the effect on student average grades. As well, the method of analysis raised questions concerning the reasonableness of pooling calculations across subject areas. On this last item, the general conclusion was that pooling was reasonable, but that, in the case of the two subjects with standard deviations quite different from the pooled value, conclusions drawn for all subjects might not apply without some further consideration.

On the main questions, first, about 40% of grades show a discrepancy of more than five marks, using the estimation procedures described. Second, from Table 4, small classes have no more than their proportionate share of deviations greater than plus or minus five. However, of the deviations greater than ten in either direction, three-quarters occur in classes smaller than twenty, even though these classes represent only about one-quarter of the total individual grades. Thus, for really extreme values of D, class size seems to play an important role. Third, Table 6 demonstrates that, when we allow more tolerance for the random error of small classes, the variation in class standards can be shown to be more than expected due to chance, especially in the larger classes. Finally, Table 7 shows that there is some consistency within a school on the standard set, so that a reasonable conclusion would be that the effect is partly teacher specific, and partly related to the school. Whether either of these cases can be labelled "error" rather than a matter of judgment was discussed.

Policy recommendations are beyond the scope of this report, as we have not considered the role of the school as an agent for social change. It is arguable that schools in economically depressed rural areas might be serving their pupils best by inflating their marks, so that they will at least have a better chance for admission to post-secondary institutions. To those who deliberately hold marks down so as to "keep up standards," perhaps the point should be made that standards and grades are not the same thing, and that standards are set when you teach, grades when you mark.

Since these results suggest that something in the order of 5% of graduates are in a position to be treated unjustly, similar studies, perhaps involving standardized testing on a large scale, should be conducted. The importance of the problem, of course, depends to a great extent on the importance of a student's grades in determining his future.

NOTE

1 I wish to thank Bob Crocker, Steve Norris, Bill Spain and an anonymous reviewer for helpful comments, and Helen Banfield for assistance with the analysis. I am also grateful to the Newfoundland Department of Education for access to data.

REFERENCES

Bull, N. L., Colbourne, W., Crocker, R. K., & Spain, W. H. *Shared evaluation: An assessment.* St. John's, Newfoundland: Institute for Educational Research and Development, Research Report, Memorial University of Newfoundland, 1977.

Glass, G. V., & Stanley, J. C. *Statistical methods in education and psychology.* Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Nagy, P., Drost, D., & Banfield, H. The Influence of Community Isolation on School Achievement. A paper presented to the Annual Meeting of the Canadian Educational Researchers' Association, Vancouver, June, 1983.

Diese Studie behandelt das Problem der unterschiedlichen Normen, die bei der Verleihung des Schulabschlusses in verschiedenen Höheren Schulen zur Anwendung kommen, und die denkbaren Ungerechtigkeiten, die Schülern aus diesen Unterschieden erwachsen. Die Reichweite des Problems wird anhand von Endnoten für das Jahr 1980 an Abgangsklassen von Höheren Schulen in Newfoundland, wo ein System der Gemeinbewertung durch Examen der Landesschulbehörde und der Höheren Schulen Nachforschungen solcher Art ermöglicht, untersucht. Der Bericht konzentriert sich auf die folgenden Fragen: (a) wie weit reicht die Erscheinung? (b) besteht ein Zusammenhang mit der Größenordnung der Schule? (c) in wie weit ist die Erscheinung auf unabsichtliche Fehler zurückzuführen? und (d) sind unterschiedliche Normen der Bewertung auf die Wirksamkeit der Schulen oder auf die der einzelnen Lehrer zurückzuführen? Weiterhin werden drei zusätzliche Themen untersucht: 1. die Hypothese einer Methode der Erfassung der Lehrer-Toleranz; 2. die Annahmen und Voraussetzungen dieser Analyse; und 3. die Unparteilichkeit der Verbesserung der Noten auf der Basis eines Toleranz-Indexes für Lehrer/Schüler Noten.

Este estudio trata sobre las normas diferentes que existen para que los estudiantes se graduen de las escuelas secundarias. Se examina la magnitud del problema usando los datos de las classes que se graduaron en 1980 de las escuelas secundarias de Newfoundland, donde existe un sistema de evaluación conjunta por medio de examenes provinciales y escuelas secundarias. El informe examina: (1) cuán amplio es este fenómeno; (b) si el fenómeno está relacionado al tamaño de la escuela; (c) que proporción del fenómeno es atribuible a errores al hazar; y (d) si las discrepancias en las normas son principalmente función de las escuelas o del profesor en particular. Asimismo, se examinan tres temas: uno, las presunciones de generosidad del método de evaluación del profesor; dos, las presunciones de este análisis; y tres, la justicia del ajustamiento de las notas en base a un indice de generosidad del profesor al marcar.

Philip Nagy, The Ontario Institute for Studies in Education, Midwestern Centre, Victoria School, 25 Joseph Street, Kitchener, Ontario N2G 1H9.