ASSESSMENT OF TEACHERS' GRADING PRACTISES

by

WILLIAM MONTE DAUNCEY
B.Ed., University of Victoria, 1976

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES
(Department of Educational Psychology)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
October, 1986

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of EDUCATIONAL PSYCHOLOGY

The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3

Date OCTOBER 11, 1986

Abstract

Educators have been using the letter grade system of grading and reporting student achievement for several decades. Since its inception, teachers have derived letter grades from a variety of grading techniques. As a result, this approach to grading has often received criticism from those who question its reliability and usefulness. The purpose of this study was to determine if letter grades could be made more reliable by statistically balancing raw achievement scores prior to aggregation for reporting purposes.

Many authors have written to attack letter grades while others have written to defend their use. Some have written to suggest alternatives to letter grades while still other writers have suggested methods of improving grading techniques. However, literature searches have shown that very little research has been done to assess teachers' grading practises and the grades they award students.

This investigative study was designed to evaluate the grading methods used by 37 randomly selected elementary school teachers. Information on their methods of grading was collected in three ways: (a) by way of a questionnaire, (b) by having the subjects weight, total, and rank a hypothetical set of raw achievement scores, and (c) by having the subjects submit class records for one reporting period from two subject

areas, mathematics and social studies. The raw scores for each class were statistically balanced and the teachers regraded their students based on the revised aggregate totals. In order to control for extraneous and subjective factors, the same grading criterion was used for both the original and the revised aggregate scores.

The rankings of the original aggregate scores were compared to those derived from the balanced aggregate scores using the Spearman Rank Correlation Coefficient. The correlations were found to be significant for each record sheet, indicating that the null hypothesis should be rejected for each of the 56 classes studied.

Analysis of the class record sheets and questionnaire responses also revealed:

(a) that 46% of the 1,314 students involved in the study received a change in letter grade in spite of the significant rank correlation coefficients. This suggests that for many students, the assignment of letter grades was unreliable and based on factors other than total score rankings.

(b) that only 5% of the respondents used methods that would apply the desired weighting factors to the raw scores. This suggests that many teachers used unreliable methods to weight assignment scores.

(c) that none of the subjects in this study used reliable methods to compensate a student who has missed one or more assignments or tests. This suggests that students who were

absent may have been unjustly rewarded or penalized when their aggregate scores were calculated.

(d) that 76% of the respondents showed a desire to learn more about collating raw scores and assigning letter grades to aggregate scores.

These results suggested that in-service instruction and pre-service training in particular aspects of grading and reporting would be justified for many members of the research sample. Areas of greatest need are those concerning the weighting of raw scores, the allocation of letter grades, and the calculation of compensation scores for students who have missed assignments.

Research Supervisor: Dr. H. Ratzlaff

TABLE OF CONTENTS

LIST OF TABLES

Acknowledgement

I would like to express my gratitude to my research supervisor, Dr. H. Ratzlaff, and to committee members Dr. W. Boldt and Dr. W. Szetela for their help and counsel.

I wish to extend my sincere thanks and appreciation to my wife, Roxanne, for her continuous support and encouragement throughout the course of my studies.

Finally, to my children, Denise and Douglas, a special note of gratitude for their patience and understanding.

CHAPTER 1

Scope of the Study

Background Information

For more than fifty years, the letter grade system has been used in education to report and document the level of student achievement. However, even after decades of use, grading practises within the letter grade system still tend to vary among teachers, departments, schools, and school districts. Many teachers acquire grading practises informally; from a colleague, department head, or by using methods that seem intuitively reasonable. Others may have had a fleeting exposure to an "acceptable" grading procedure during teacher training.

> It is surprising that many teachers embark on their careers in the schools without ever having been taught anything about grading. This very important part of the teacher's task is often ignored in teacher training. Perhaps it is just taken for granted, or perhaps it seems too intractable to be taught. (Hills, 1981, p. 316)

Consequently, when inconsistencies in grading practises exist, equal levels of achievement are not necessarily given equal letter grades. For example, a grade of "C" in one classroom may be a "C-" or a "C+" in another. As a result, grades and grading practises become somewhat suspect and often lose credibility with the post-secondary institutions, prospective employers, students, and other individuals or agencies that

depend upon reliable academic assessments as a basis for decision making. This notion is supported by Hills (1981) who states,

> Faulty grading can be seriously detrimental to students. It can give them misleading information and interfere with the learning process. It can give misleading information for others who legitimately use school grades as part of their procedures in evaluating people. And it can mislead those who would study the instructional process to improve it or to hold it accountable to the society that supports it. (p. 317)

However, for most students unreliable calculations of a term's grades will not cause much harm. For example, if a "C" is awarded instead of a C+ or a C- , the differential effects upon the student will not vary significantly. Discrepancies such as these will seldom lead to incorrect decisions as C+, C, and C- are all symbols of "satisfactory" achievement. On the other hand, when a student has been given an F when he merited a passing grade, some far reaching consequences could result. Such an error could cost the student credit for the course, or even an extra year in school if a course failure prevents promotion. At the other end of the scale, a C+ instead of a B, or a B instead of an A could result in the loss of a scholarship, a bursary, or an opportunity to attend a particularly competitive post-secondary institution. Of course, the inverse is equally significant. For example, if a

student is erroneously awarded a passing grade, he will be given credit for a course when it is not warranted and when he is not sufficiently prepared to go on to the next level.

In elementary schools, marks and other measures of student achievement form the basis for determining whether a student will be promoted into the next grade, whether he will receive an academic award or whether he will be a candidate for learning assistance or a special remedial class. In junior high school, term grades are used as in the elementary school setting but with the additional use of identifying honour roll students. Senior high schools use term grades not only for promotion and retention but also to determine which students will be required to write final exams, which students will be eligible for scholarships and bursaries, and which students should be encouraged to attend post-secondary institutions.

Although there are those who disagree with the letter system of grading and reporting (see Curwin, 1978), those who do agree feel it may be one of the teacher's most important roles. "One of the most important professional responsibilities of academic staff is that of allocating and reporting marks (or grades) as measures of student performance" (Isaacs and Imrie, 1981, p. 3). Curwin suggests that teachers should question themselves as to the reliability of grades in fulfilling the reporting and learning functions. He states, "The one basic requirement of the report is that all information received must have the same meaning that was

sent by the grader. Grades must be valid, reliable and have a high degree of predictability. But they don't."
(Curwin, 1978, p. 60).

Identification of the Problem

It seems clear that there is a need for teachers' grading practises to be as reliable and accurate as possible. Most school boards also recognize the need for reliability in evaluation and reporting. As stated in the Administrative Handbook of one large district,

> The Board of Trustees believes that evaluation and reporting of pupil achievement are integral and vital parts of teaching and learning. The Board further believes that evaluation and reporting should be based on reliable information about pupil performance in relation to the specific requirements of a program, course, or grade. (School District Administrative Handbook, 1978)

With reliable and accurate reporting, consistency among schools and classrooms is maintained, credibility in the eyes of the community is preserved, and educational decisions can be made more confidently and correctly. However, the question remains: Are the teachers' current grading practises as reliable and accurate as possible? Curwin (1978), for one, claims they are not.

With the introduction of the letter grade report cards in British Columbia's elementary schools, there is the assumption that teachers will have the background knowledge and the resources necessary to collate scores, grade, and report student achievement reliably. As indicated previously, many teachers may not have the necessary understanding or training to perform such tasks well. Many teachers simply do not know how to combine student achievement scores in such a way as to yield reliable letter grades for reporting purposes.

> Teachers agree that giving and reporting marks (or grades) is one of their most uncomfortable responsibilities. Some lack confidence in the marks they assign; others believe their marks are fair, but find them difficult to defend. Behind these negative attitudes lies the fact that the basis for assigning marks is often unclear. (Ahmann and Glock, 1981, p. 417)

As a result, many teachers combine and calculate scores in a manner that seems intuitively reasonable, and by virtue of their lack of knowledge, blissfully accept their letter grades as being reliable assessments of their students' work. Many assume that by using raw scores their term assignments will be appropriately weighted when combining marks for their term totals. Many teachers seem to regard the mark value of each assignment as being an indicator of the weight contributing to the total term score (ie. an assignment worth 100 marks contributes twice as much toward term totals as does an assignment worth 50 marks). It would seem that one of the

more common methods of collating achievement results for grading purposes involves simple totalling of raw scores or averaging (highest term total or average receives an A etc.). In extreme cases, where standard deviations among term assignments are very different, simple summations and averaging can result in highly erroneous aggregate scores and unreliable reporting, as will be demonstrated in a subsequent example. Also, some teachers still abide by the percentage 'rule-of-thumb' in terms of test, assignment, or term results. They may, for example, consider any score above 80 per cent to be equivalent to an 'A', even in extreme cases where an 80 per cent is the lowest score or where the highest score is less than 80 per cent. Grades obtained from these and other similarly questionable methods are accepted as being credible and reliable by interested parties (such as parents, principals, prospective employers, and higher educational institutions) and are used as a basis for making important decisions. In short, teachers often have considered neither mean scores nor the variability of scores when evaluating pupil achievement. As will be shown, calculating term totals from several assignments may yield unexpected and unreliable results if the class averages and the amount of variation in each assignment are not considered. "If we hope to maintain the weighting scheme originally chosen, we must take into consideration the differences in variability. A failure to do this will result in inequities" (Ahmann and Glock, 1981, p. 426).

Before examining why mean scores and standard deviations should be employed, a review of how final or term totals are typically obtained is in order. Final course grades are typically based on various types of student achievement, such as short quizzes, term projects, laboratory work, and examinations. Thus, assigning course grades involves combining these various measures to obtain a composite or term score in such a way that each assignment receives its intended weight. For example, if it has been decided that the final examination should count twice as much as the short quizzes, then the grades should reflect this emphasis. Similarly, if the laboratory work is to make up 25 per cent of the final course grade, then it is important that it will be represented to that extent. It is not uncommon for teachers to communicate this type of information to their students at the beginning of the course. Students are particularly interested in knowing which elements will be included in the final grade and the relative weight each will be given.

Determining how much weight should be given to each of the various assignments is a matter of judgment. This judgment is guided by the importance of the various instructional objectives, the teaching emphasis given to each type of course activity, and other similar considerations. The teacher must decide on the composition of the grades given in each course and the relative emphasis to be allocated to each component. Without assigning desired weightings to assignments and by using only the natural weightings, it is

possible for a relatively minor assignment to have a very strong, if not dominant, influence on the course grade. Also, the natural weightings may reflect a totally different emphasis on the various instructional objectives than was actually used by the teacher during the course.

When the decision has been made concerning what proportion of the final aggregate score is to be allocated to each measure of student performance, the measures must be combined in such a way that the desired weighting is obtained. This brings us back to the importance of calculating class means and standard deviations. To illustrate the problems of the various grading practises in the weighting of achievement data, which is the problem of interest in this study, consider a simple example (taken from Gronlund, 1974).

Let us assume that we want to combine scores on an examination and scores on a term project, and that we want them to contribute equally to the final (or term) grade. Suppose the range of scores on the two measures are:

Range of scores on examination	20 to 120
Range of scores on term project	40 to 60

If we simply add together the examination score and the project score for each individual, the final grade will be determined largely by how well the students performed on the examination. To demonstrate this, compare a student who is highest on the examination and lowest on the term project

(Student A) with a student who is highest on the term project and lowest on the examination (Student B).

	Student A	Student B
Examination score	120	20
Term project score	40	60
Composite score	160	80

It is obvious that simply adding the two scores will not give them equal representation in the composite score. Nor will calculating an average score provide equal weighting. The examination score has much greater influence than the term project score.

In a situation such as this, teachers frequently attempt to equate the influence of the two measures by making the maximum possible score for both assignments equivalent. In the above example, this would mean multiplying the students' term project score by two to make the total possible score 120 for both measures of achievement. The following shows what happens when this is done:

	Student A	Student B
Examination score	120	20
Term project score (x2)	80	120
Composite score	200	140

As the example demonstrates, equating on the basis of maximum possible score does not provide for equal emphasis in the composite score either. The examination score still has the greater influence when the two are combined. This is because the contribution a given measure makes to a composite score depends largely on the variability, or spread, of the scores in the set.

To equate the examination score and the term project score, the variation of the scores in each set must be made equal. This is done by adjusting the standard deviation of one of the sets of scores so that it is equal to the other set. The method most often used is the standard T-score transformation. This changes each set so that the class average (mean) is 50 and the standard deviation is 10. This is not an easy task without the use of a computer or a sophisticated calculator, a great deal of time, and some statistical knowledge. However, once the scores are 'balanced', weightings can be applied to each transformed score with the confidence that the desired weightings will be obtained. In this simple example, the two assignments are weighted equally. Our example would then look like this:

	Student A	Student B
Examination T-score (weighted)	100	80
Term project T-score (weighted)	80	100
Composite T-score (weighted)	180	180

Clearly, the examination and the term project scores are equal; both students receive the same composite score. This is possible only by equating the two assignments in terms of means and standard deviations. In this case, the standard T-score was used for both exercises.

The main questions to be addressed then are the following:

1. Is there a significant difference between the grades reported by teachers using intuitive methods and those reported when means and standard deviations are utilized and applied? In other words, if teachers were to follow the advice of many authors and employ basic statistics when aggregating marks, would the results be significantly different from those obtained from existing grading practises?

2. Are the raw achievement scores combined and weighted reliably?

3. Are the methods used to compensate students who have missed assignments or tests fair and reliable? Are stuents given grades that penalize or reward them for being absent for a test or assignment?

4. Would any improvement in results be of such a magnitude to justify the time and expense required to retrain teachers?

It was the purpose of this study to investigate these questions to determine if teachers' aggregating and grading methods could be improved.

DEFINITION OF TERMS

The following are terms used in this study:

1. ASSIGNMENT SCORES: "Assignments" refer to the various exercises, tests, laboratory work, term projects, etcetera, that a teacher might assign his students during the course of instruction in a particular subject. An "assignment score" is the result that each student would receive after completing a particular assignment. At the end of each term, a total or "term score" is obtained for each student by summing their respective results on each assignment. For reporting purposes each student's term score is then ranked largest to smallest and letter grades are assigned based on the rankings of the total scores.

2. BALANCED SCORES: "Balanced Scores" or "statistically balanced scores" refer to assignment scores that have undergone a linear transformation to alter the mean and standard deviation. When several class assignments have been "balanced", they have been given equal means and equal standard deviations. Balanced assignments contribute equally towards the final course mark.

3. FAIR: For the purpose of this study, "fair" is used to mean "just" or "not favouring one more than the other". Fair grades are those that do not penalize or reward individual students more than they do other students. Fair grades are also consistent. For example, two students having

the same raw scores and subjected to identical grading criteria would be expected to receive identical grades. However, if one of the students were to receive a higher grade than the other, then the grades would not be consistent and would be considered "unfair".

4. LETTER GRADE: For the purpose of this study, a letter grade is defined as: "1) a summary symbol, 2) evaluating a substantial segment of achievement which is 3) attained by a pupil in a course and 4) assigned by a teacher 5) for the purpose of record and report" (Ministry of Education, Science and Technology, 1979, p. 9). A letter grade should be reserved for judgements that include evidence available on student achievement in a complete course or a segment of a course such as a reporting period. Letter grades are usually derived from the evidence available in a set of scores for each pupil. It is also assumed that a teacher has a systematic procedure for assigning grades. Because a letter grade is "used to record and report", it is assumed that a letter grade is a "permanent record usually retained for a long period of time in the school record system and used to communicate with individual students, parents, teachers, and other educational authorities and potential employers" (Ministry of Education, Science and Technology, 1979, p. 10).

5. MEAN: "Mean" is synonymous with "arithmetic mean" or "average" and is a measure of central tendency. It is calculated by summing a series of numbers and dividing the sum by the number of numbers summed. The value of the mean is

affected by the individual values of all of the scores in the set of data. In this study, the mean is denoted as X_m and individual raw scores as X_i.

6. RELIABLE: Reliable grades are those that are "dependable and fit to be relied upon". Grades that are reliable provide an accurate indication of a student's performance and are worthy of being used as a basis for decision making.

7. T-SCORES: "T-Scores" are a linear transformation of standard scores. Standardized T-scores are transformed to a set of data having a mean of 50 and standard deviation of 10. Calculation of T-scores is achieved by multiplying a standard score (z score) by 10 and adding 50 to the product. The shape of the distribution is not changed with the transformed scores.

8. STANDARD DEVIATION: "Standard Deviation" is a measure of dispersion, scatter, heterogeneity, or variation in a set of data. A set of scores with great heterogeneity will have some large deviation scores ($X_i - X_m$). Standard deviation is calculated by first summing the squared deviation scores and then dividing the sum by the number of scores. The square root of the quotient is the standard deviation. In this study, the standard deviation is denoted as SD.

9. STANDARD SCORES: "Standard Scores" can be used to describe the position of a score in a set of scores by measuring its deviation from the mean of all scores in standard deviation units. A standard score (called a z-score)

is obtained by subtracting the mean from an observed score and dividing the difference by the standard deviation. The mean and standard deviation of the transformed raw scores (z-scores) are 0 and 1 respectively. In this way, then, the difference between an individual score and the mean score of the group, is scaled in terms of standard deviation units. This has the advantage in that all deviation scores are measured in the same units.

10. TERM: A "term" is an instructional period where courses can be offered in whole or in part. The end of each term is normally marked by the evaluation and reporting of student achievement in each of the courses offered during that period. It is stipulated by the British Columbia Ministry of Education that, for schools in British Columbia, there will be at least 3 reporting periods (or terms) during each school year.

11. WEIGHTING: "Weighting" refers to the amount of influence a particular exercise or assignment has in determining the total score for that subject. For example, an exercise might contribute 20 per cent toward the total mark for the course, meaning that 20 per cent of the total course mark is derived from that particular assignment. Weighting should normally reflect the importance or the emphasis placed on the objectives covered in the assignment. "NATURAL WEIGHTING" refers to the contribution an assignment gives to the course mark using only raw score data. In other words, for a given set of assignments in a particular course, each

assignment will contribute a certain percentage of the total score. The contribution will be determined by the variability of the raw scores. This raw score influence is called "natural weighting".

Research Questions

Specifically, the questions of interest are:

1. Is there a significant difference between the ranking obtained from a teacher's course record and the ranking obtained from statistically balancing this same record?

2. Do a teacher's course assignments contribute the intended portion of marks towards the course total? (Are the term totals, from which the letter grades are derived, collated and weighted reliably?)

3. Are teachers' grades reliable and fair for students missing one or more assignments or tests? Are students compensated fairly and reliably when they have been absent and have not received a score for one or more assignments or tests?

4. Would the revised results be of such a magnitude to justify the time and expense required to retrain teachers and revise university courses and/or programs?

In this study, the questions of interest relate to the

methods used by teachers to collate achievement data for reporting purposes. The hypothesis to be tested is that teachers, for one reason or another, do not collate achievement scores reliably. It is assumed that teachers do not consider the calculation of class means or standard deviations when combining achievement scores. As a result, the assignment scores are contributing disproportionate weights toward the course totals. It is hypothesized that the method most commonly used by teachers to combine course marks is simple summation of the raw scores followed by ranking each student's total. It is also hypothesized that the summation method of ranking will be significantly different from rankings obtained from statistically balanced assignment scores.

Rationale

Prior to the introduction of the letter grade report card for elementary schools in 1979, elementary school report cards were primarily anecdotal, and achievement symbols (Needs Improvement, Satisfatory, Good), were based largely upon subjective evaluation. When the format of the report was changed to comparative achievement and letter grades were used to reflect this achievement, teachers were given very little in-service training on the new reporting procedure. The British Columbia Ministry of Education did publish booklets

("Construction and Uses of Classroom Tests: A Resource Book For Teachers" and "Grading Practices:Issues and Alternatives") to assist teachers in their evaluation of student achievement. Although these booklets were made available, it appears that few teachers studied them in detail prior to using the new report cards.

Consequently, it would seem that few teachers are aware of the importance of the mean and standard deviation when aggregating assignment scores. Most teachers use only the raw scores total for determining the ranking of their students. Many educators seem to follow the erroneous belief that the assignment with the largest possible score should contribute the greatest percentage toward the final course mark.

The absence of statistical analysis and 'balancing' in the past may have been partially due to a lack of training and partially due to a lack of time and resources. A statistical calculator is almost a necessity but a great deal of teacher time is still required to calculate the standardized scores - time that some might find difficult to justify in light of the many other professional commitments of teachers. However, with the advent of the micro-computer, "mark-keeping" programs can be devised to take the time consumption and drudgery out of the statistical balancing of student assignment scores. Even so, many teachers seem to abide by their previously established grading practises. They are not aware of the need to consider the effects of different variances within their assignment scores, and therefore are oblivious to the need for

improved grading practises. Hills (1981) supports this notion and states, "The standard deviation . . . is what most teachers would not think about without some training" (p. 319). Perhaps if there is a perceived need for a change in the way scores are collated, teachers will be receptive to an alternate and improved method of combining students' scores for reporting purposes.

It is hoped that the outcome of this study will help to determine if a more effective method of combining achievement scores is needed and whether more in-service and pre-service training is required to improve the grading practises of teachers.

Delimitations

For the purposes of this study, the sample was selected at random from teachers of grades four to seven level who teach in a large metropolitan area of central British Columbia. Examples of teachers' records were selected from two subject areas: mathematics and social studies. The instruments used by the sample subjects to collect student achievement data were not evaluated. It was assumed that all instruments were reliable and valid. We were interested only in the combined total score for the term and the assigned letter grade for each student as determined from the collation of term or course assignment scores.

This study was not concerned with the validity of the

educational objectives for each of the two subjects sampled. It should be recognized that within the curriculum stipulated by the Provincial Ministry of Education, considerable flexibility is provided to elementary teachers. Educators may select from a variety of content, media, pedagogical techniques, and whatever else is necessary, to fullfil the demands of the curriculum and the needs of the students in the most effective manner. Consequently, the course content may vary among teachers of the same subject. Units and unit objectives may also be taught with different emphases. This flexibility nullifies any opportunity to "standardize" elementary subjects in terms of units, unit objectives, assignments, and assignment emphasis on course marks. As a result, the validity, emphasis, and appropriate weightings of the objectives were outside the scope of this study. They were assumed to be appropriate for the grade level, the students, and the curriculum.

Justification of the Study

Why is this issue important in education? Firstly, student achievement should be reported as reliably and accurately as possible. "Faulty grading can be seriously detrimental to students. It can give them misleading information and interfere with the learning process" (Hills, 1981, p. 281). Secondly, parents, prospective employers, and

the students themselves view report card grades as accurate indicators of student achievement and use the information as a basis for decision-making and for evaluating people. Thirdly, unreliable or inaccurate grades may lead to erroneous decisions regarding bursaries, scholarships, diplomas, or other such awards based on grades. More importantly, improper assessment of student performance may result in incorrect decisions regarding a student's success in a particular course. In addition, inaccurate grading "can mislead those who would study the instructional process to improve it or to hold it accountable to the society that supports it" (Hills, p. 281). A fifth reason why this issue is important to education is due to the fact that there are many misconceptions and confusions among educators as to what constitutes an effective grading procedure. Hills also supports this perception and states, "there are many irrational ideas about grading that seem sound until they are considered in detail. To avoid some of these irrational notions, a teacher must learn to recognize what is wrong with them" (p. 280).

The implications of this study pertain to teacher pre-service and in-service training. If the samples of teacher grading practises are significantly inferior when compared to the statistically balanced and transformed scores (T-scores), then it may be prudent to examine the pre-service training teachers receive in evaluating student performance

and achievement. Is this an area that needs more attention in the pre-service training of teachers? Should universities make courses in evaluation mandatory for student-teachers? Also, if it is apparent that grading practises can be improved, should school districts develop in-service programs to enhance teacher grading practises? These are important questions if the results show that grading practises can be improved. Many authors reflect this importance by suggesting that teachers should strive to improve their grading procedures. Hills (1981), for example, states,

> Grading is a complicated problem. There are many pitfalls; erroneous practises abound; and the teacher cannot safely proceed on the assumption that what was done when she was a pupil is satisfactory for the students now in her care. Good teachers will try to do a better job than was done for them as pupils, rather than simply carrying forward the mistakes of the past. (p. 280)

On the other hand, if the present grading practises are not significantly inferior and are in fact generally as good as the statistically balanced grades, then we can conclude that teachers, similar to those in the sample, are evaluating and grading student achievement effectively. If this is the case, then universities and school districts need not concern themselves with revising their present measurement and student evaluation components or programs. The implication would be

that the training teachers receive and the evaluation techniques teachers employ are both adequate and reliable.

CHAPTER 2

Survey of the Literature

From the review of the literature, there appear to be very few empirical or investigative studies done to assess the grading practises of teachers. Although there is an abundance of theoretical information on how teachers should evaluate and report the achievement of their students, it seems that no one has actually tried to find out if teachers are putting theory into practise. Hills (1981) reinforces this fact when he states,

> There is little empirical research on grading. Although some facts are well-established, there has been far too little study of such an important topic. Therefore, everyone seems to feel entitled to an opinion, and no data exist to refute most of them. (p. 317)

Several authors, however, have suggested innovative methods of grading for special educational situations or for particular courses. For example, David Cohen (1973) discusses a means of improving the evaluation of the Australian science curriculum. Gensley (1969) provides an insight into alternate methods of evaluating "gifted children". Many other writers advocate alternate forms of evaluation for their particular subject area or for their particular specialty situation. However, none of the authors offer any insight into the adequacy of the grading practises of "regular" classroom teachers.

Other writers either attack or defend the use of letter

grades as a symbol of student achievement. For example, W.J. Stewart (1975) condemns letter grades and professes a new system for evaluating the "multi-faceted academic-personal-social growth of elementary school students" (p. 174). He states that one grade does not represent adequately the complexity in a child's academic, social, and emotional growth. Instead, he recommends a multi-dimensional system that would accommodate (a) a wide range of goals, (b) uniqueness of each student's abilities and needs, (c) an account for each child's need to develop a positive self-concept, and (d) a need to provide parents with more useful and more meaningful information regarding a child's progress. Stewart claims that his evaluation-reporting system would de-emphasize skill and subject matter mastery but would stress a child's ability to use skills and subject matter for dealing with life's daily problems. His multi-dimensional system would focus only upon individuals and would not stress comparisons with other class members as do letter grades. While Stewart points out what may be considered by some to be weaknesses of the letter grade system, he does not address what teachers are presently doing to cope with the problems he identifies. He cites no research to justify his claim that a new system of "evaluation-reporting" is required.

Marshall (1971) also criticizes the letter grade system. He claims that grades (or any "codes which refer to rank, grade, or position on a scale") tend to be very subjective. He states, "Grades survive primarily because they provide

disguise. Almost anything can be, and is, read into them. Letter grades . . . are general terms, in themselves meaningless" (p. 350). Marshall indicates the benefits of an "individual descriptive or nongrading" system of evaluation and reporting that would include providing teachers with "more time, greater freedom, and new interests in teaching" (p. 352). He also states that "teachers are employed to teach, not to grade or psychoanalyze" (p. 352). He concludes by stating, "A teacher should want, be allowed, and be asked to say what he or she means, in words specifically pertinent to a given inquiry or situation" (p. 353). As with Stewart, Marshall makes no mention of teachers' current grading practises.

A third criticism of the letter grade system is stated by Richard Curwin (1978). He claims that "grades don't work" because they lack accuracy, consistency, validity, reliability, and predictability. He offers four alternatives (Self-Grading, Contract Grading, Peer Grading, and Blanket Grading) to "reduce the dangers of the grading system" (p. 61). Although he states that "research studies clearly indicate that teachers grade the same students' work differently" (p. 60), he does not elaborate on the findings or the nature of the research. He ends by stating that the alternatives "undoubtedly will be an improvement over traditional grading systems" (p. 64).

Defenders of letter grades as symbols of student achievement take a stance opposite to the authors mentioned

above. Robert L. Ebel (1974), for example, defends grades against the criticisms often aimed at grading. He claims that grading can be improved. "The remedy is not to get rid of grades but simply to do a better job of grading" (p. 3). He indicates that all alternatives to the grade are less cost efficient and less informative than the "familiar summary statistic, the grade."

Although Ebel cites no research to demonstrate how classroom teachers collect achievement information upon which to base the assignment of grades, he does elaborate on how grades can be improved. He, like many other grade supporters, advocates calculating the standard deviations when aggregating and weighting student assignment scores. By using a sample of student marks, he shows how equating the variation of the various assignments can lead to improved grading. "Thus the influence of one component on a composite depends not on total points or mean score but on score variability" (Ebel, 1972, p. 350). Many other grade defenders also support this notion. Hills (1981), for example, states, "The key element in weighting variables that are to be combined is the standard deviation of their values" (p. 320). Like Ebel, Hills also uses examples to show how calculating the standard deviation is essential for accurate grading. With conversion charts, he demonstrates how the standard deviation can be estimated without the use of a hand calculator. Hills summarizes by stating, "Using these procedures will help to extinguish the criticism that grades are meaningless" (p. 328).

Gronlund (1981) also supports the use of letter grades in classrooms. He states, "When letter grades are supplemented by other methods of reporting, these grades themselves become more meaningful" (p. 520). He too utilizes an example to demonstrate how grading can be improved by making mean scores and standard deviations equal when combining and weighting raw scores. He also expresses the importance of considering the dispersion of scores. "Thus, to properly weight the components in a composite score, the variability of the scores must be taken into account" (p. 522). However, in his example, Gronlund shows how the variability in each set of assignment scores can be equated using three different methods: (a) by using the range of scores, (b) by converting all sets of scores to stanines, and (c), the most refined system, by using standard deviations.

Many other authors, (such as Ahmann & Glock, 1981; Townsend & Burke, 1975; Guilford & Fruchter, 1973) also support the use of letter grades as a method of grading and reporting student achievement. But, while all profess the importance of equalizing the variability in a set of scores, none cite any research that would indicate whether or not teachers collate and weight scores properly. None indicate whether or not practising teachers actually consider the variation in their scores. Perhaps by improving the methods of grading as suggested by the defending authors the controversy surrounding the usefulness of the letter grade as a symbol of student achievement will diminish.

Grading Systems

Over the last fifty years, several attempts have been made to replace the letter-grade evaluation system with other marking systems or to abolish school marks altogether. Curwin (1978) has outlined some of the possible alternatives to evaluation systems in schools. Although these alternatives have not been particularly popular in the past, they are mentioned here to reveal the options available to teachers and to indicate the various strengths and weaknesses of each evaluation system.

The first alternative Curwin offers is called "Self-Grading". In this system, the teacher supplies the criteria and the data for the students to use in assigning their own grades. As a variant to self grading, the students can also generate their own evaluation criteria, as long as the criteria remain within the guidelines of the course. Curwin states that the main benefit of this approach is that students are encouraged to evaluate their own work and thereby learn about themselves and their abilities. However, critics may question the reliability of the grades resulting from such a system. It would seem that the reliability is dependent on the suitability, appropriateness, and validity of the evaluation criteria as well as on the responsibility and maturity of the students.

Another option available to teachers is called "Contract Grading". Although similar to self-grading, this system is

more structured. It centers around a contract that is agreeable to both student and teacher. The contract stipulates the work to be accomplished and the time frame within which the work must be completed. The four elements of the contract are: (1) criteria for successful completion of the contract, (2) data or results of work that indicate how well the criteria have been met, (3) an analysis of or comparison between the data and criteria, and (4) grade (or points) earned. A disadvantage of contract grading is that "the contracts sometimes become inflexible. Students might set unrealistic goals--too high or too low" (Curwin, 1978, p. 62). Another difficulty is that students in such a system tend to do only what is stipulated in the contract and to avoid the spontaneous extension of their learning beyond what is written.

Curwin describes "Peer Grading" as another alternative to the letter grade system of evaluation. This method utilizes a contract formulated by a group of students. Upon completion of the contract, the group evaluates itself and each member receives the same group grade. The groups can be formed on the basis of a subject area, a homogeneous ability grouping, or around a mutual interest. Curwin claims,

> Students often are inspired to learn as energy is generated through working with friends and peers on a common project, but a negative attitude by some students can drag the rest of the group down. If some students work harder than others, they may perceive it as unfair

to receive the same grade as those who worked less hard. (p. 62)

A fourth alternative to grading is referred to as "Blanket Grading". With this system, "each student in a class receives the same grade (usually an A or B) providing that some minimum standard is met. Those students who do not reach the specified standard must negotiate a grade individually with the teacher" (p. 64). Unlike Curwin´s other alternatives, Blanket Grading requires a continuous evaluation process as part of the regular class activities. This can become a negative feature of this alternative since a very systematic and well organized evaluation system is required. Simply declaring that everyone will receive an A or B without any other form of evaluation would prove to be counter-productive; students would be less motivated to achieve to their potential. Another negative aspect of Blanket Grading is that administrators often have a difficult time accepting this system of grading, especially when a disproportionate number of A´s or B´s are given. This, according to Curwin, can cause the teacher "personal or professional harm" (p. 64). Curwin also emphasizes that this approach to grading needs to be very thoroughly explained to the administration and to the students before being used.

In reference to these four alternate grading schemes, Curwin states, "If you use any of the alternatives presented here, you will avoid many (although not all) of the problems

intrinsic to grading, but you still must meet some of the original purposes of grades -- especially the function of supplying information to parents" (p. 64).

Of the more popular grading systems, most are variations on one of three approaches: (a) percent grading, (b) criterion-referenced grading, and (c) norm-referenced grading. Each has been popular in the past as a method of determining grades in the educational setting.

Percent Grading, which is currently the least popular of the three basic approaches, involves scores that are averaged and converted to a percent. With this method, the percent itself may represent the grade. For example, mathematics might be reported on a report card as simply "Mathematics 73%". Another method of reporting percent grading is to convert a percentage to a letter grade. An A, for example, may represent scores ranging from 80% to 100%.

Percent Grading has not gained popularity mainly because of the obscurity of the grades. A score of 80%, for example, tells little more than 80 percent of the questions were answered correctly. Such a grade does not reveal how much a student learned in relation to intended outcomes nor does it reveal how he compared to the achievement of other students. With such grades it is difficult to determine if the 80 percent represents exceptional, average, or poor achievement. If it is the highest score, then it is a very good score. If it is near the mean or median score, then it is only average,

and if it is one of the lowest scores, then it represents poor achievement. It is impossible to tell what level of achievement 80 percent actually represents.

Although percent grading was once a popular method of grading, it has lost its appeal in favor of the other two major systems, namely the criterion-referenced and norm-referenced approaches to grading.

Criterion-referenced grading

Criterion-referenced grading involves expressing a student's achievement in relation to some prespecified standards or criteria. These standards are usually concerned with the level of mastery to be achieved by the students and are often stated in terms of: (a) the specific tasks to be performed, and (b) the percentage of correct answers to be obtained on a test that measures a clearly defined set of learning objectives. As a result, these standards are absolute and are not relative to other students or to previous individual achievement. All students who achieve at the same level or master the same objectives achieve the same grade. "No student is failed simply because he or she achieved lower than other students provided the student achieves the criterion" (Gray, 1980, p. 490). Grading in the criterion-referenced system is simpler than many other grading methods. Those who reach the specified criterion on each learning objective receive credit, and those who do not, fail.

Failure to satisfy the criterion may also necessitate that the student repeat the required tasks until the objectives have been achieved. Consequently, nearly all students in the criterion-referenced system pass.

The criterion-referenced grading technique with the mastery learning approach, makes it relatively easy for teachers to identify problem areas in their curriculum and to identify weaknesses within individual students. Criterion-referenced tests, for example, will reveal which students had difficulty (or failed to meet criterion) with which objectives. Those objectives that were poorly grasped by a majority of students and therefore need to be retaught will also be revealed. "Thus, deficiencies are quickly corrected and students do not get farther and farther behind " (Gray, p. 490). Students in this system are permitted the opportunity to progress at their own levels - some will meet the criterion sooner than others. However, by the end of the course or term, most students will have demonstrated an acceptable level of performance on the required learning objectives.

Gronlund (1981) states, "To effectively use absolute level of achievement as a basis for grading requires that (1) the domain of learning tasks be clearly defined, (2) the standards of performance be clearly specified and justified, and (3) the measures of pupil achievement be criterion referenced" (p. 524). While these conditions are relatively easy to satisfy in the mastery approach, they become very

difficult to apply to nonmastery learning objectives. This prompts Gronlund to state, "The criterion-referenced system of grading is much more complex than it first appears" (p. 524). Because most educational programs in the elementary school setting will include both mastery and nonmastery objectives, the elementary teacher will have considerable difficulty setting the criteria for acceptable performance with all objectives. Even when the criteria are established, Gray (1980) claims that "it is difficult to defend such standards to parents, even if they have been set by curriculum experts in a curriculum guide, because they are based primarily on judgment" (p. 491). Gronlund (1981) claims that these judgments are "likely to be contaminated by achievements to some unknown degree. Thus, the lack of reliablility in judging achievement in relation to potential, and in judging degree of improvement, would result in grades of low dependability" (p. 524). This is very significant since demonstrated improvement and learning potential have been widely used as a basis for grading the nonmastery components of the criterion-referenced grading system in elementary schools. It is also interesting to note that when reporting student achievement on nonmastery objectives, rank order based on scores or on number of objectives completed is often used to provide some measure of relative position within the class. Rank and relative position are terms that connote the norm-referenced approach to grading student achievement.

Norm-referenced grading

Traditionally norm-referenced grading has been the most common approach to grading in education. It involves rank ordering students and expressing each student's achievement in relation to the achievement of his class-mates. Unlike criterion-referenced grading, a norm-referenced grade does not indicate what a student has actually achieved; it only describes the relative position within the class.

In the extreme form of norm-referenced grading, the class serves as the normative group. Assignments and norm-referenced tests are designed to produce a wide range of scores and it is this variablility that is used to determine the distribution of final grades. For example, if a student ranks high in the group, he will receive a high grade. If his relative achievement is low, he will receive a low grade. Since the grading is based on relative achievement, the students' grades are influenced by both his performance and by the performance of the group. Consequently, a particular student will do better in a low achieving class than in a high achieving class. Because the norm-referenced approach assumes a normal distribution in the population of students, most will receive a grade indicating average achievement. A few students will receive A's and a few will fail. However, this is more characteristic of normal curve grading, which is one of several variations of the norm-referenced technique available to teachers.

Normal curve grading is the extreme form of

norm-referenced grading. In this system, a fixed percentage of students receive each grade. A commonly used grade distribution in this system assigns the top 5% an A, the next 24% a B, the middle 40% a C, the next 24% a D, and the lowest 7% a failing grade (Gray, 1980, p. 488). A weakness with this approach is that some students must receive failing grades regardless of their levels of achievement. For example, a student who achieves a score of 70% may receive a failing grade if his score is one of the lowest in the group. On the other hand, if 70% represented a high score, then the student would receive a high grade. Another weakness is that the same grade does not necessarily represent the same level of achievement. In one class, a student's achievement may yield a B; in another class, the same achievement might be equivalent to a C or a D. Many stories have been told how non-academics, derelicts, and in one American air force story, the students' wives were persuaded to enroll in courses where normal curve grading was used. The theory was that these "class fillers" would absorb the failing grades and thereby allow the rest of the students to pass the course. These stories highlight the weaknesses of normal curve grading.

Another variation to norm-referenced grading is "Pass-Fail grading" (Gray, 1980, p. 488). This system has only two possible grades; either Pass or Fail. Any student who achieves at a level that is at or above the minimum acceptable standard, receives the "Pass" grade. The rationale behind this type of grading is that such a system takes the

pressure off students and encourages them to take courses they might normally avoid due to fear of receiving a lower grade. However, studies have shown that students tend to be less motivated to do their best in courses using the pass-fail approach. "Thus, pass-fail grading does not fulfill any of the purposes of grades--communication, motivation, and prediction. . . . In general its use should be discouraged" (Gray, 1980, p. 489).

A more satisfactory variation to norm-referenced grading is "standard score grading". This technique involves converting raw scores from each test or assignment to standardized z scores, or T scores, averaging the T scores and assigning grades based on the average T score value. For example, average T scores of 65 and above might receive an A and scores between 55 and 64 might receive a B and so on. The main advantage of this grading system is that arithmetic operations can be performed on the T scores. This is an important feature because "we can meaningfully combine scores in a way that adjusts appropriately for the fact that the various contributing tests had different degrees of variability" (Gray, 1980, p. 489). Another advantage to standard score grading is that teachers can weight assignments differentially. Seldom are all tests and assignments considered to be of equal importance when combining marks. Standard score grading allows teachers to assign a desired weight to each of the contributing assignments before combining the scores. However, "Good judgment has to enter

into the grading process. . . . The point to keep in mind is that norm-referenced grading simply rank orders the students; it indicates nothing about actual level of achievement" (Gray, p. 490).

One of the major criticisms of the extreme form of norm-referenced grading is the fact that a fixed percentage of students receive each grade. As stated previously, this implies that a small percentage of students will always receive failing grades, regardless of the levels of achievement. However, to overcome the negative aspects, the British Columbia Ministry of Education (1979) offers another variation to norm-referenced grading. It suggests the use of standard score grading but with a grade distribution based on student ability rather than on a normal curve. The methods of establishing the grade distribution makes this approach different from the other methods of norm-referenced grading already mentioned.

Essentially, an ability-based grading distribution means that the grade distribution for a given class reflects the ability levels of the students in that particular class. "The proportion of different grades assigned is predetermined by the ability levels within the group" (British Columbia Ministry of Education, p. 26). Therefore, it is possible that, for a particular class, a larger number of grades may be given at one level than in any other. Some distributions may have levels where no grades are awarded. To create such an ability-based distribution requires the estimation of the

general ability level of students within the class. Normally, the necessary information can be obtained from standardized achievement test scores or from general ability tests, or both. Information on student ability can also be obtained from past performance records, such as previous report cards or school records. Using this information, a grading distribution can be determined for each class. It is important to point out that the grades or percentile rankings assigned to individual students during the estimation process are not to be considered as expectation levels for the students so categorized. For example, students who receive A's may not be the same students who scored above the 96 percentile on the standardized test.

While the use of T scores maintains the advantages of standard score grading, a grade distribution based on student ability overcomes the problems of predetermined failures. These methods, combined with the use of "natural breaks" in the distribution of collated scores to determine letter grade cut-off points, "offers a very satisfactory approach to the assignment of letter grades" (British Columbia Ministry of Education, p. 27). This variant to the norm-referenced grading procedure has been adopted by the school district involved in this study.

There are several reasons why the norm-referenced approach to grading has remained one of the most popular methods of evaluating and grading student achievement. Firstly, grades based on relative achievement appear to be

more readily understood by parents, administrators, prospective employers, and others who have an interest in student achievement. Relative position in a group or class, as indicated by the letter grade, can be comprehended quickly and easily. Very little information needs to accompany the letter grade to explain its meaning; due to its frequent use and familiarity it is almost common knowledge that an A represents outstanding work while a D or E indicates that the student has experienced considerable difficulty. Other grading systems require checklists or grading criteria to provide meaning to the letter grades. These lists need to be interpreted and understood before full value can be derived from the associated letter grade.

Secondly, parents generally are interested in knowing how well their child performed relative to the others in the class. The competitive nature of people, the same sort that drives them to "keep up with the Jones", emerges when discussing their child's academic achievement. For example, they might want to know if their child's A was a high A or a low A or how many others in the class received an A. People often appear to use relative position as an indicator of success, whether it be in business, industry, society, or education. The further up the ranking scale a person is the more successful he is perceived to be.

A third reason why norm-referenced grading methods have been more common is due to the predictive qualities of this grading system. Many of the uses of grades are predictive in

nature. For example, employers want to be able to predict the abilities and success of the students they hire. Colleges and universities use grades to predict the success of their applicants. "If all students received nothing but pass-fail grades, it would be very difficult for college admissions personnel, for example, to select from among applicants since all applicants would have the same grades" (Gray, 1980, p. 489). Because the norm-referenced system attemps to produce maximum variability of scores, the correlation coefficient is enhanced and becomes more useful for prediction purposes than coefficients derived from other grading systems.

A fourth reason why norm-referenced grading has maintained its popularity is due to the relative ease of setting standards and grade distributions. For educational institutions using the extreme form of norm-referenced grading, the normal curve is used to establish the grade distribution. In this case, the standards and grade distributions between various classes or courses remain similar regardless of course content or student ability levels. Consequently, the norm-referenced grading system is easier to administer and to apply. Once the final score ranking has been achieved, the number of grades to be allotted at each level can be easily determined. Even the variations of norm-referenced grading, those that do not strictly adhere to the normal curve distribution, are more easily administered than most of the other grading techniques. One of the main disadvantages of the criterion-referenced systems, for

example, is the difficulty of formulating the criterion and standards for acceptable performance. With a norm-referenced approach, these tasks are much simpler; the administration of grades is more mechanical and mathematical, requiring less time and judgment from the teacher.

Another reason for the popularity of norm-referenced approaches is due to the motivational aspects. One of the main purposes of grades is to motivate students to perform as close to their potential as possible. Norm-referenced grading has been shown to motivate students more effectively than do most other methods of grading. For example, criterion-referenced approaches to grading create less motivation in students. Gray (1980) states, "if all courses were pass-fail, student motivation (and subsequent achievement) would decrease across the board" (p. 492). Norm-referenced systems can capitalize on the competitive nature of students and motivate them to do their best.

In British Columbia, the Ministry of Education stipulated that all public schools would report comparative achievement with a seven level letter grade system in grades four to twelve. This decision was based on the results of a survey conducted by the Ministry in the late 1970s. This method of reporting necessitated the use of some form of norm-referenced grading procedure within the school system. Booklets were published by the Ministry to provide guidelines and to assist teachers in utilizing norm-referenced techniques in their

classrooms. The reporting and grading methods used in the elementary schools of the school district involved in this study follow these guidelines and are outlined in chapter 3 of this study.

CHAPTER 3

Methodology

It was the purpose of this study to evaluate the grading methods used by 37 randomly selected teachers. Information on their methods of grading was collected in three ways: (a) by way of a questionnaire, (b) by having the subjects weight, total, and rank a hypothetical set of raw achievement scores, and (c) by having the subjects submit their mathematics and social studies class record sheets for one reporting period.

Sample and Population

The sample was selected at random from a list of male and female teachers who teach at the grades four to seven level in a large metropolitan area of central British Columbia.

The accessible population from which the sample was selected consisted of teachers who teach at the intermediate level in elementary schools. At this level, teachers customarily teach all of the Ministry prescribed subjects to one class. The sizes of most elementary school classes normally range from twenty five to thirty students. Teachers are responsible for administering the prescribed curricula, and for measuring, evaluating, grading, and reporting student achievement. In the elementary school setting, measurement instruments consist primarily of teacher made tests and assignments.

Questionnaire

A questionnaire was given to each member of the sample of teachers in order to gain some insight into their teaching experience, their knowledge of collating achievement scores for grading purposes, and their knowledge of basic statistics (particularly standard deviation). A Likert-type scale was used for some questionnaire items and the closed form format was used to obtain the demographic and grading procedure information (Appendix A). Such questions as the following were asked:

1. My colleagues at this school take report card marks seriously.

2. Letter grades are an effective method of informing parents of their child's progress and achievement.

3. To the best of my knowledge, parents are generally satisfied with the letter grade system of reporting.

4. I feel very comfortable with the letter grades I award students.

5. I am confident that the letter grades I assign are accurate and reliable.

6. I would like to learn more about collating and assigning letter grades to raw scores.

7. More in-service sessions on grading and reporting should be offered.

8. Universities should offer more pre-service instruction in effective grading and reporting.

9. The report card format allows sufficient information

to be communicated to parents, students, etc.

10. In the space provided, briefly describe how you combine assignment and test scores to determine letter grades.

11. In reference to the method described in #10, which other subjects are also graded this way? (all subjects, mathematics, language arts, science, social studies, art, P.E., music, French)

12. In the space provided, briefly describe how you arrive at a final term score for a student who has been absent for one or more tests or assignments. In other words, how do you compensate for a legitimately missed assignment?

The questionnaire also collected data using the completion format. Such questions as the following were asked:

1. Which is your age category?
(<25,25-30,31-40,41-50,51-60,>60) (years)

2. How many years teaching experience do you have? (combine both public and private)
(0-2,3-5,6-10,11-15,16-20,>20)

3. How many courses in statistics or measurement and evaluation do you have? (disregard unit value) (0,1,2,3,>3)

4. What is your teaching assignment? (full time, part time)

5. If your answer to #4 was "part time", explain your teaching assignment (percentage of time, subjects, etc.)

Included with the questionnaire was a sample of five fictitious achievement scores for five fictitious students and the weight that each assignment contributed to the final course score. The respondents were asked to collate and rank the scores using the method most familiar to them. This served as a graphic demonstration of the collating methods commonly used by teachers.

Fictitious Data

The hypothetical data consisted of student names, their raw scores on each of the fictitious assignments, and the weight each assignment was to contribute to the total term score. The teachers were asked to weight, collate, and rank the assignment scores using their accustomed methods. Subsequently, these rankings were analyzed and compared to the rankings of the same data after the raw assignment scores had been statistically balanced.

For comparison purposes, the raw scores on each of the fictitious assignments were transformed to T-scores and balanced prior to being weighted and collated. Each student's revised total was then ranked and compared to the teachers' ranking of the raw score data. By analyzing the results of

the fictitious data, it was possible to determine the potential for error and lack of reliability when teachers collate student achievement scores. An example of the hypothetical data supplied to the subjects is listed in Table I. The bracketted information (weighted total score) was not supplied to the respondents and is listed here only to indicate the distribution of the raw scores. Table II displays the same data after it has been balanced so that each assignment now has a mean of 50 and a standard deviation of 10. Statistical balancing was applied before the assignment scores were weighted and collated. In this study, the ranking of the "Weighted Total Scores" was the main point of interest.

Of particular interest in this example is how the student rankings in Table II differ from those in Table I. This demonstrates how, even with only five students, statistical balancing can have a noticeable effect on student rankings. With a larger number of students, the effect may be even more pronounced. Since the rankings are used to determine letter grades, it is imperative that the collated scores be as reliable as possible.

Record Sheets

Examples of teachers' records were selected from two subject areas: mathematics and social studies. Each of the two record sheets obtained from the respondents revealed the students' names, the raw scores for the various assignments and tests, the students' collated scores and letter grades, as

NAME	ASSIGNMENTS raw scores 1	2	3	4	5	TOTAL SCORE	RANK	WEIGHTED TOTAL SCORE	RANK
1. Student A	50	55	65	57	95	[64.4]	[3]	[72.7]	[1]
2. Student B	69	60	58	80	83	[70.0]	[1]	[71.7]	[2]
3. Student C	45	72	52	71	75	[63.0]	[4.5]	[66.4]	[3]
4. Student D	75	95	45	83	25	[64.6]	[2]	[53.8]	[4]
5. Student E	73	83	60	63	36	[63.0]	[4.5]	[45.8]	[5]
WEIGHTS (%)	10	20	20	10	40				
[X_m]	[62]	[73]	[56]	[71]	[63]				
[SD]	[14]	[16]	[8]	[11]	[31]				

Table I - Fictitious data to be collated and ranked

NAME	ASSIGNMENTS T-scores 1	2	3	4	5	TOTAL SCORE	RANK	WEIGHTED TOTAL SCORE	RANK
1. Student A	41	39	61	37	60	47.6	4	51.8	2
2. Student B	55	42	53	58	56	52.8	1	52.7	1
3. Student C	38	49	45	50	54	47.2	5	49.2	3
4. Student D	59	64	36	61	38	51.6	2	47.2	5
5. Student E	58	56	55	43	41	50.6	3	48.7	4
WEIGHTS (%)	10	20	20	10	40				
X_m	50	50	50	50	50				
SD	10	10	10	10	10				

Table II - Fictitious data after being statistically balanced

well as the weighting that each assignment was to contribute to the collated scores.

At the end of a term (just prior to the distribution of report cards), copies of the record sheets for both subject areas were collected from the respondents. Each class record sheet was subjected to the following procedure. With the aid of a computer, the assignment raw scores were statistically balanced and converted to T-scores. They were then weighted, collated and ranked utilizing the same assignment weightings as used by the teacher. The revised totals and rankings were returned to the original respondent to be regraded. It was hoped that, by having the respondent regrade the balanced results, all procedural and subjective factors affecting the students' grades would be held constant for both the original and the balanced sets of data.

The differences in the rankings of the raw score and the transformed scores were analyzed in two ways. Firstly, the rankings were compared to determine if there were statistically significant differences in the correlation between the two rankings. Secondly, letter grades resulting from both the raw score and the T-score rankings were compared to determine how many students received a change in letter grade. The results from these analyses were compiled and tabulated. A complete documentation of the results can be found in chapters four and five of this study.

Grading Policy

In order to promote consistency in the allocation of letter grades in the elementary school setting, the school district used in this study has in effect a policy outlining how the letter grade frequency distribution should be determined for each class. Essentially, the policy states that the letter grade distribution should be based on student ability estimates derived from the national percentiles of standardized achievement test results. This is achieved by administering the standardized Canadian Achievement Test (CAT) to the intermediate grades once each school year. The resulting national percentiles for each class are used to assist in the establishment of an approximate grade distribution for that class. Table III illustrates the guidelines that are supplied to teachers in order to assist them in relating national percentiles to letter grade distributions. If, for example, in a given class 3 students scored between 96 and 100 on the CAT national percentile, 7 between 76 and 95, 4 between 61 and 75, 9 between 41 and 60, 4 between 26 and 40, 3 between 6 and 25, and none between 0 and 5, then approximately three students should receive an A, approximately 7 should receive a B, and so on. The district guidelines suggest only an approximate distribution as teachers are urged to consider also the standards of acceptable achievement for their particular grade when allocating letter grades. However, if the guidelines were to be strictly followed, then the three students with the highest

NATIONAL PERCENTILE	LETTER GRADE
96 - 100	A
76 - 95	B
61 - 75	C+
41 - 60	C
26 - 40	C-
6 - 25	D
0 - 5	E

Table III - Letter grade distribution guidelines

cumulative scores from the various term assignments would receive the A´s, while the next seven highest would receive the B´s, etcetera. This procedure "takes into consideration the ability levels of the pupils in a particular grade or class" (British Columbia Ministry of Education, 1979, p. 26). However, the school district policy also states, "In considering the cumulative marks for each subject, "natural breaks" were used further to establish letter grade distributions" (School District Policies and Regulations). This approach is also advocated by the British Columbia Ministry of Education in their booklet on grading practises. "Natural breaks" refer to the gaps that appear in the ranked distribution of students´ cumulative scores. The gaps can be used as grade cut-off points to separate one grade category

from another. Combining the natural breaks and the group ability methods to determine the letter grade distribution allows "teachers to make judicious decisions as to what constitutes high and low achievement" (British Columbia Ministry of Education, 1979, p. 26). Using these procedures to establish the letter grade distribution for each class highlights the importance of combining assignment scores reliably.

Data Collection and Analysis

The teachers were given record sheets at the beginning of the reporting period on which to record the achievement scores for both academic subjects. At the end of the term the record sheets and the questionnaires were collected and analyzed. Information from the questionnaires was collated and tabulated. The achievement scores for each of the two subjects were statistically balanced and aggregated, and the students were ranked based on the aggregated score. This "balanced" ranking was correlated with the teachers' ranking using the Spearman Rank Correlation Coefficient r_{ranks}.

Additionally, the results from the regrading of the balanced data were compared to the grades assigned using the teachers' original data. The number of students who received a change in letter grade was recorded and documented. The direction and magnitude of the change was also tabulated.

From the tabulated results, it was possible to

generalize as to whether teachers, similar to those in the sample, could benefit from in-service and pre-service training on revised grading techniques. It also became apparent, from the results, whether or not grading practises can be and should be made more reliable.

CHAPTER 4

Results

The data collected in the study were divided into four categories: (a) demographic information concerning the sample, (b) data concerning the subjects' attitudes and perceptions toward grading and reporting, (c) analysis of the sample's grading techniques, and (d) analysis of the subjects' course record sheets.

Sample

A sample of 37 subjects was used in the study. While all 37 responded to the first data request (copies of their record sheets), 34 responded to the second set of data (survey and regraded record sheets). Three of the 34 subjects elected to withhold their regraded record sheets and submitted only the completed questionnaire. A total of 31 subjects submitted both regraded record sheets and completed questionnaires. Of the 31 respondents, 25 provided regraded data for both mathematics and social studies. The sample provided raw scores and letter grades for 1,314 elementary school students from 56 classes. Tables IV through IX provide the frequency distributions of the sample's demographic variables. Tables X through XVII display the frequency distributions of the subjects' responses to the attitude and perception questions pertaining to grading and reporting.

Demographic Variables

a) Gender

Male	24
Female	8
Total	34

Table IV - Distribution by Gender

b) Age (years)

Less than 25	0
25-30	0
31-40	25
41-50	8
51-60	1
Greater than 60	0
Total	34

Table V - Distribution by Age (years)

c) Years of university

two	0
three	1
four	10
B.Ed	15
MA or MEd	7
other (ie diploma).	1
Total	34

Table VI - Distribution by University Training (years)

d) Number of measurement and/or evaluation courses taken

none	11
one	21
two	1
three	1
More than three	0
Total	34

Table VII - Distribution by Measurement Courses Taken

e) Teaching assignment

part time	3
full time	31
Total	34

Table VIII - Distribution by Teaching Assignment

f) Teaching Experience (years)

0-5	0
6-10	7
11-15	15
16-20	9
Greater than 20	3
Total	34

Table IX - Distribution by Teaching Experience (years)

Of the 34 subjects completing the second set of data, most (74%) were male teachers and most (74%) were in the 31 to 40 year age category. Twenty four per cent of the respondents were in the 41 to 50 year age category. Teachers who had completed a Bachelor's degree made up 44% of the sample while 29% had four years university experience and 21% had completed a Master's degree. Nearly all of the respondents (91%) were full time teachers and 44 per cent had between 11 and 15 years teaching experience. Twenty six per cent of the teachers had 16 to 20 years teaching experience. Many of the subjects (62%) had completed one measurement and/or evaluation course while 32% had not taken any measurement or evaluation courses.

In general, the typical respondent was a 31 to 40 year old male teacher who was employed full time and who had between 11 and 15 years of teaching experience. He typically held a Bachelor's degree and had completed one measurement and/or evaluation course.

Attitudes and Perceptions Toward Grading and Reporting

Tables X through XVII reflect the sample's attitude and perception towards grading and reporting student achievement. The subjects were to respond on a five-point scale the extent of agreement between the feeling expressed in each statement and their own personal feeling. The five options were: Strongly Disagree (SD), Disagree (D), Undecided (U), Agree (A), and Strongly Agree (SA).

g) My colleagues at this school take report card marks very seriously.

SD.....	0
D	0
U	1
A	15
SA.....	18
Total	34

Table X - Distribution of colleagues´ attitudes toward report card marks

(h) Letter grades are an effective method of informing parents of their child´s progress and achievement.

SD.....	1
D	5
U	2
A	22
SA.....	4
Total.....	34

Table XI - Distribution of attitude toward letter grade effectiveness

(i) To the best of my knowledge, parents are generally satisfied with the letter grade system of reporting.

SD	0
D	2
U	1
A	28
SA	3
Total	34

Table XII - Distribution of agreement regarding parents' satisfaction toward letter grade reporting systems.

(j) I am very confident that the letter grades I assign are accurate and reliable.

SD	0
D	0
U	4
A	23
SA	7
Total	34

Table XIII - Distribution of agreement regarding subjects' own grading ability

(k) I would like to learn more about collating and assigning letter grades to raw scores.

SD.....	0
D	5
U	3
A	16
SA.....	10
Total.....	34

Table XIV - Distribution of agreement regarding subjects' desire to learn more about marking and grading

(l) More in-serivce sessions on grading and reporting should be offered.

SD.....	0
D	5
U	4
A	11
SA.....	14
Total.....	34

Table XV - Distribution of agreement on evaluation of in-service sessions

(m) Universities should offer more pre-service instruction in effective grading and reporting.

SD.....	0
D	3
U	4
A	14
SA.....	13
Total.....	34

Table XVI - Distribution of agreement regarding subjects' opinion on pre-service training

(n) The report card format allows sufficient information to be communicated to parents, students, etc.

SD.....	0
D	8
U	6
A	16
SA.....	4
Total.....	34

Table XVII - Distribution of agreement regarding effectiveness of the report card format as a communication device

Nearly all (97%) of the teachers surveyed agreed that their colleagues treat marking and grading seriously (44% Agree;53% Strongly Agree). A majority of the respondents (71%) agreed that letter grades are an effective method of informing parents of their child's progress and achievement. Fifty nine per cent of the sample also agreed with the feeling that the report card format allows sufficient information to be communicated to parents, students, and other interested groups. However, 18% were undecided and 24% disagreed with this statement. In response to the statement concerning parent satisfaction, most respondents (91%) agreed that parents were satisfied with the letter grade system of reporting.

In response to the questions pertaining to teachers' grading and reporting techniques, a majority of the subjects

(88%) agreed (68% Agree; 20% Strongly Agree) that the letter grades they assign are reliable and accurate. However, most respondents indicated a desire to learn more about collating and grading raw achievement scores (76% Agree or Strongly Agree; 9% Undecided; 15% Disagree). The percentages of respondents agreeing to the need for more in-service and pre-service instruction in this area were 74% and 79% respectively. On the same questions, 12% were undecided for both in-service and pre-service sessions while the percentages of respondents who disagreed with the need for in-service and pre-service instruction were 15% and 9% respectively.

In general, the typical respondent was one who (a) thought his colleagues took report card marks seriously, (b) thought letter grades are an effective method of informing parents of their students progress, (c) thought that parents were satisfied with the letter grade system of marking, (d) thought the British Columbia Ministry of Education report card format allows sufficient information to be communicated to parents and students, (e) thought that more in-service and pre-service sessions on evaluation and grading should be offered by school districts and universities, and (f) displayed an interest in learning more about aggregating and grading raw scores.

Grading Techniques

The record sheets and the responses to the grading and reporting section of the survey were analyzed to determine the

grading procedures used by the subjects. Although all of the respondents employed a variation of the norm-referenced approach to grading, the sample's grading techniques were classified further according to the following grading methods:

1. Method One - Summing raw scores and converting each student's total to a percentage prior to grading by natural breaks (see page 53).

2. Method Two - Summing each student's raw scores and converting the total to a percentage. Letter grades are assigned based on the numerical value of the percentage as well as a subjective factor. Examples of subjective factors were: (a) adjusting letter grades to reflect individual effort; (b) adjusting letter grades to reflect a change in a student's performance from one term to another; or (c) adjusting letter grades to compensate for missed assignments.

3. Method Three - Summing weighted raw scores for each student and converting the sum to a percentage prior to grading according to natural breaks.

4. Method Four - Summing each student's raw scores and converting the total to a percentage. Letter grades are assigned based on the numerical value of the percentage. This method is commonly called "percent grading" (see page 32).

5. Method Five - Term assignments, projects and tests are given letter grades rather than numerical scores. Each student's final grade is the average of all the letter grades awarded that student during the course. For averaging purposes, each letter grade is given an ordinal value.

6. Method Six - Raw scores are balanced to have equal means and equal standard deviations before being weighted and collated. The collated scores are ranked and graded according to natural breaks.

Table XVIII gives the frequency breakdown of each grading method. Table XIX displays how extensively each respondent applied their grading technique.

Method One	24
Method Two	5
Method Three	2
Method Four	2
Method Five	2
Method Six	2
Total	37

Table XVIII - Distribution of grading method

Every school subject	9
Core subjects (Math, Science Social Studies, English)	24
Indiscriminate use	1
Total	34

Table XIX - Distribution of application of grading technique

The respondents were also asked to describe how they determined a final score for those students who have missed one or more tests or assignments. Their responses were classified under the following methods:

1. Method One - Calculate the average percentage on only the assignments completed and ignore the missed assignments.

2. Method Two - Estimate a score for the missed assignment by looking at: (a) the absent student's previous work, (b) the marks other students of similar ability obtained on the missed assignment.

3. Method Three - Assign a mark of zero for the missed assignment.

4. Method Four - Calculate a letter grade based on previous work and reduce by one letter grade for missed assignment(s).

5. Method Five - Calculate an average T score based on all completed work and assign that value to the missed assignment(s). When all assignment scores have been statistically balanced this method essentially maintains rank position and does not penalize nor reward the student for being absent.

Table XX shows the frequency of the methods used by the sample to arrive at a mark for an absent student.

Method One	25
Method Two	5
Method Three	2
Method Four	1
Method Five	0
Other	1
Total	34

Table XX - Distribution of compensation method used

The results of the procedures used to compensate students for missed assignments (Table XX) indicated that most respondents (74%) calculated an average score based solely on the work the absent student had completed. The data also reveals that 15% of the respondents use subjective means to estimate a score for a missed assignment.

The participants in the study were asked to apply their grading technique to a set of fictitious data (see page 48) and rank the five students from highest to lowest. Table XXI shows the frequency of the ranks as determined by the various grading techniques.

In response to the second research question (page 16) regarding the reliability of the aggregation and weighting of raw scores for grading purposes, the sample's record sheets and the questionnaire responses pertaining to grading techniques were analyzed. The analysis revealed a high degree of similarity between the aggregation methods evident on the

Student A B C D E	
1,2,3,5,4	16
3,1,4.5,2,4.5	6
2,1,3,5,4	3
1,2,4,5,3	2
4.5,1,3,2,4.5	1
2,1,5,3,4	1
2,3,5,1,4	1
1,2.5,2.5,5,4	1
2,1,4.5,3,4.5	1
Incomplete	2
Total	34

Table XXI - Distribution of ranks of fictitious data

record sheets and the aggregation techniques described by the respondents on the questionnaire. However, the grading methods used in response to the fictitious data section of the survey were less consistent with the methods evident on the grade sheets and in the written descriptions. The most frequent discrepancy concerned the weighting factors to be applied to the fictitious data. Although most respondents did not normally weight raw scores, many attempted to apply weights to the fictitious data.

As shown in table XXI, 18% of the sample obtained a rank order for the fictitious total scores that was consistent with the most common grading method identified in this study (see

Table XVIII). Nine per cent of the respondents applied aggregation and weighting techniques that are advocated by many experts (see page 28). The suggested methods are tabulated as Method 6 in Table XVIII. Many subjects (47%) used aggregation and weighting techniques that were inconsistent with any of the other grading methods previously identified in this study. Their methods essentially involved combining weighted raw scores. However, they obtained results that were very similar to both the ranks of the aggregated weighted raw scores (Table I) and to the ranks of the aggregated weighted balanced scores (Table II).

The results of the classification and tabulation of the various grading methods used by the respondents, as tabled in Table XVIII, revealed that a 65% of the sample employed an aggregation technique that involved summing the raw scores and converting the collated scores to a percentage prior to grading by natural breaks. The results also showed that 14% of the sample used an aggregation technique that involved summing the raw scores, converting the combined scores to a percentage, and using the numerical value of the percentage as well as subjective factors to determine the letter grade. Only two of the 37 respondents (5%) used "reliable" methods to apply the desired weighting factors to the raw scores.

Record Sheets.

The rankings of the statistically balanced totals were compared to the rankings of the raw score totals using

Spearman's Rank-Correlation Coefficient. Table XXII lists the subject number, the number of students in each of the subject's two classes (n), the rank correlation coefficient (r_s) , and the critical values of Spearman's rank Correlation Coefficient for testing the null hypothesis of no correlation with a two-tailed test at the alpha=.01 level of significance (Glass, G., Stanley, J., 1970, p. 539). Table XXIII displays the distribution of the Spearman Rank Correlation Coefficients.

Of the 56 correlations computed, all had r_s values that were greater than the corresponding critical value. These results indicate that the correlations between the rankings of the raw score totals and the rankings of the balanced score totals were positive and significant. Therefore, the null hypothesis (H_o:rho=0) should be rejected at the .01 level of significance in favour of the alternate hypothesis for each of the 56 record sheets analyzed.

In the second set of data, the subjects were asked to regrade their students based on the statistically balanced totals provided. The letter grades based on balanced scores were compared to the letter grades assigned initially by the respondent. The differences between the two letter grades were classified according to magnitude and direction of "change". For example, a change from an initial letter grade of B to a balanced score letter grade of C+ was recorded as a one-letter grade decrease. A change from a C+ to a balanced letter grade of A was recorded as a two-letter grade increase. Table XXIV

SUBJECT #	MATHEMATICS R_s	n	CRITICAL VALUE	SOCIAL STUDIES R_s	n	CRITICAL VALUE
001	.941	24	.537	.976	24	.537
002	.873	34	<.478	.957	34	<.478
003	.955	18	.625	.985	18	.625
004	.983	24	.537	N/A	N/A	N/A
005	.919	33	<.478	.984	33	<.478
006	.954	14	.716	N/A	N/A	N/A
007	.953	27	.505	.975	27	.505
008	.975	14	.716	.970	14	.716
009	N/A	N/A	N/A	.968	21	.576
010	.785	25	.526	.903	25	.526
011	.992	24	.537	.943	33	<.478
012	.746	20	.591	.857	20	.591
013	.987	16	.666	.981	36	<.478
014	N/A	N/A	N/A	.980	24	.537
015	.984	17	.645	.939	17	.654
016	N/A	N/A	N/A	.949	18	.625
017	.937	16	.666	.797	16	.666
018	.992	28	.496	.962	28	.496
019	.956	13	.745	.995	13	.745
020	.955	30	.478	.960	32	<.478
021	.977	24	.537	N/A	N/A	N/A
022	.979	33	<.478	.876	32	<.478
023	.988	32	<.478	.957	31	<.478
024	.821	16	.666	.900	16	.666
025	.931	23	.549	.947	23	.549
026	.984	29	.487	.976	27	.505
027	.929	16	.666	.918	16	.666
028	.986	24	.537	.902	24	.537
029	.986	28	.496	.869	28	.496
030	.972	25	.526	.960	25	.526
031	.999	16	.666	.875	16	.666
Total		643			671	

Table XXII - Spearman Rank Correlation Coefficients (R_s) Relating Raw Score and Balanced Score Totals For (a) Mathematics and (b) Social Studies

	Frequency	
Score Interval	Math	S.S.
0.970 - 1.000	14	9
0.940 - 0.969	6	9
0.910 - 0.939	4	2
0.880 - 0.909	0	3
0.850 - 0.879	1	4
0.820 - 0.849	1	0
0.790 - 0.819	0	1
0.760 - 0.789	1	0
0.730 - 0.759	1	0
0.700 - 0.729	0	0
Total	28	28

Table XXIII - Distribution of Spearman Rank Correlation Coefficients

shows the frequency, the magnitude, and the direction of the letter grade changes.

For analysis purposes the range of letter grades (A to E) was divided into three categories in accordance with the Ministry of Education letter grade interpretations: Above Average, Average, and Below Average. Letter grades A and B were grouped into the Above Average category; letter grades C+, C, and C- were grouped into the Average category; and letter grades of D and E were grouped into the Below Average category.

One letter grade increase	238 (18.1%)
One letter grade decrease	313 (23.8%)
Two letter grade increase	28 (2.1%)
Two letter grade decrease	23 (1.8%)
Three letter grade increase	4 (0.3%)
Three letter grade decrease	1 (0.1%)
No change in letter grade	707 (53.8%)
Total	1314 (100.0%)

Table XXIV - Distribution of letter grade changes

The differences in letter grade were also analyzed to determine how many "changes" would result in a shift from one category to another. Table XXV shows the frequency of category shifts that resulted when the original letter grades were replaced by letter grades based on statistically balanced scores.

Above Average to Average	112 (8.5%)
Average to Above Average	54 (4.1%)
Below Average to Average	44 (3.3%)
Average to Below Average	29 (2.2%)
Letter grade change but no change in category	368 (28.1%)
No letter grade/category change	707 (53.8%)
Total	1314 (100.0%)

Table XXV - Frequency of category shift

The results of the letter grade comparison, listed in Tables XXIII, XXIV, and XXV, indicated that 46% of the 1,314 grades initially awarded to students were changed when the students were regraded on the basis of the balanced score totals. Thirty nine per cent of the revised grades (or 18% of all grades) also involved category changes.

Finally, the sample's record sheets were analyzed to determine how many letter grades were based on miscalculated raw score totals. Table XXVI displays the frequency of errors, as well as the direction and magnitude of the letter grade correction. Table XXVII shows the frequency of category changes that resulted when incorrect raw score totals were recalculated and the corresponding letter grades revised.

One letter grade increase.....	2 (0.2%)
One letter grade decrease.....	6 (0.4%)
Two letter grade increase.....	0 (0.0%)
Two letter grade decrease.....	2 (0.2%)
Corrections that did not affect letter grade.....	53 (4.0%)
No errors.....	1251 (95.2%)
Total	1314 (100.0%)

Table XXVI - Frequency of letter grade changes resulting from arithmetic errors

Above Average to Average.....	4 (0.3%)
Average to Above Average.....	1 (0.1%)
Below Average to Average.....	0 (0.0%)
Average to Below Average.....	2 (0.2%)
Letter grade change but no change in category	3 (0.2%)
No letter grade change.....	1304 (99.2%)
Total.....	1314 (100.0%)

Table XXVII - Frequency of category shift resulting from letter grade recalculation

The analysis of the record sheets revealed that about 5% of the 1,314 student grades assigned by the sample were based on incorrectly computed raw score totals. However, only about 1% of the students received incorrect letter grades as a result of the calculation errors. Of the ten improperly assigned grades, seven resulted in an incorrect category placement as well.

This concludes the results section of this study. Chapter five will contain a short summary, an analysis of the data in response to the questions of interest, an outline of the limitations, and finally, the possibilities for further research.

CHAPTER 5

Conclusions and Recommendations For Further Study

Summary

The general problem that this investigative study considered was to determine if teachers' grading methods could be revised so as to increase the reliability and fairness of the letter grades they award to students. It was difficult to adequately deal with this problem in one study. It would be almost impossible to analyze each of the many components comprised in the marking and grading processes. Consequently, the scope of this study was restricted to focus only on the effects of applying statistical balancing to raw achievement scores prior to the calculation of student term totals. From this perspective, the following questions of interest evolved:

1. Is there a significant difference between the rankings of the aggregated scores calculated by teachers using intuitive methods and the rankings of aggregated balanced scores?

2. Are the term totals, from which the letter grades are derived, aggregated and weighted reliably?

3. Are teachers' grades reliable and fair?

4. Would the revised results be of such a magnitude as to

justify the time and expense required to re-educate teachers?

The raw scores supplied by the sample in the first set of data were converted to T-scores, weighted, and collated to obtain a "statistically balanced" total score for each student listed on the record sheets. The weightings for the balanced data were made identical to those used initially by the respondents. In the second set of data, the subjects were asked to use the balanced scores to regrade their students, and also to complete the questionnaire listed in Appendix A.

Spearman Rank Correlation Coefficients were calculated to compare the rankings of the raw score totals with the rankings of the statistically balanced totals. The critical values of Spearman's Rank Correlation Coefficient for testing the null hypothesis of no correlation with a two-tailed test (alpha = .01) were obtained from Glass and Stanley (1970, p. 539) with the appropriate n values corresponding to the class sizes. The letter grades based on raw scores were compared to those based on balanced scores to determine if significantly different letter grades would result when raw scores were statistically balanced prior to being aggregated for grading purposes. The results of the comparison were recorded in terms of the change in letter grade that occurred when the balanced score totals replaced raw score totals. The grade changes were analyzed to determine how many students would have received a category change (see page 73) if their original grades were replaced with balanced score grades.

Tables XXIV and XXV display the results of these comparisons and analyses. The raw score grades were also analyzed to determine how many students received raw score grades based on miscalculated term totals. The numbers of category shifts that resulted from miscalculated grades were also recorded. The results of these analyses are tabulated in Tables XXVI and XXVII.

The questionnaire was used to collect data pertaining to the sample's demographic variables, the sample's attitude and perception of grading and reporting, and the sample's grading techniques. The results are tabulated in Tables IV through XXI in chapter four.

Research Question 1

In response to the first research question regarding the difference between the rankings of the raw score totals and the rankings of balanced score totals, Spearman Rank Correlation Coefficients were calculated for each of the subject's two record sheets. The results indicated that the rankings of the raw score totals were significantly and positively correlated with the rankings of the balanced score totals (see page 71). The values of the Spearman Rank Correlation Coefficients ranged from +.746 to +.999 (see Table XXIII). From these results it can be concluded that there is a significant relationship between the rankings of the raw score totals calculated by the subjects and the rankings of

the balanced score totals. It can also be concluded that the methods used by the subjects to aggregate raw scores for reporting purposes are reasonably reliable when compared with the statistical balancing methods advocated by many authors.

Since norm-referenced grading, the method used in British Columbia schools, bases letter grades on the rank order of the aggregate totals, significant positive correlations between the raw score totals and the balanced totals suggest there will be a significant and positive correlation between the letter grades resulting from each method of combining scores. In other words, the letter grade a student receives under the raw score method of combining scores should, in most cases, be similar to the letter grade he would receive when the balanced score technique is applied. Both methods should yield similarly distributed letter grades for a given class. It can be concluded, then, that on the whole, the students should receive reasonably reliable grades (see page 2).

However, of the 1,314 students involved in this study, 46% received a change in letter grade when their grades were based on balanced score totals rather than raw score totals (Table XXIV). This would suggest that, for many students, the assignment of grades was unreliable and was based on, or influenced by, factors other than the total score rankings. Since several respondents inquired about converting the T-scores on the record sheets to per cent scores, it is possible that the values of the converted per cent scores were

influencing factors for some subjects during the regrading process. The lower per cent scores (based on an arbitrary maximum of 275) resulting from the T-scores would also explain why more letter grades decreased than increased (see Table XXIV).

In reference to the consequences of unreliable grading practises (see page 2), the letter grade changes suggest that 46% of the students in this study received inconsistent and misleading information that could potentially lead others to formulate erroneous decisions and conclusions. In addition, 18% of the 1,314 students received letter grade changes that resulted in category shifts (Table XXV). These students are most likely to suffer the consequences of erroneous decisions as academic programs tend generally to cater to the below average, average, and above average ability groups. For example, students who have average ability could be erroneously placed in a special program designed for "below average achievers" or they could be erroneously enrolled in an enriched program for students with above average ability. In either case, the average student's academic needs may not be satisfactorily met.

In light of the current literature, the results of Spearman Rank Correlation Coefficients were somewhat surprising. Why are the rankings of the raw score totals (as calculated by the subjects) so strongly significant and positively correlated with the rankings of the balanced score

totals in this study? Many authors have supported the notion that raw scores can only be collated and weighted reliably when they have been balanced to have equal means and equal standard deviations (see page 27). For example, Ahmann and Glock (1981) state, "If we hope to maintain the weighting scheme originally chosen, we must take into consideration the differences in variability. A failure to do this will result in inequities" (p. 426). But these and many other authors have failed to mention the importance of the rank correlation for "between" scores. From the results of this study, it appears that the rank correlation coefficient for between scores can influence the rank order of the aggregated scores. Indeed, the "inequities" that result when aggregating raw scores with large differences in standard deviations can be reduced or eliminated with a large positive rank correlation for between scores. However, these inequities can also be amplified by a large negative rank correlation. Such is the case with the example taken from Gronlund (1974) on page 8 of this study. Gronlund has used the extreme situation (as have other authors) of large variability and a very large negative between scores rank correlation ($r_s = -1$) to emphasize the effects of combining scores with different standard deviations. In their examples, student A scores very high on the first assignment and very low on the second while student B scores low on the first and high on the second. But, this is seldom the case with scores taken from teachers' record sheets. More commonly, students tend to maintain a relatively

stable rank position within the class. Above average students tend to obtain the above average scores on most assignments while the below average students tend to achieve the lower scores. Therefore, the rankings for between raw scores would more likely be positive. It is interesting to note that elementary teachers question assignments that result in "unusual" class rankings. For example, some subjects questioned the validity of the fifth assignment in the fictitious data (see Appendix A) because the rankings on that assignment failed to fit the ranking pattern established by the other four assignments. Students who placed high in the ranking on the first four assignments should place relatively high in the ranking of the fifth assignment also. Although a rank correlation of $r_s=+1$ would be extremely difficult to obtain in a class of twenty or thirty students, the stability of rank position suggests there might be a reasonably strong positive rank correlation for between scores in many class record sheets. Considering the results of this study and the different standard deviation values in the raw score data, the likely presence of positive rank correlations for between scores may have reduced the effect of variance on the sample's aggregate scores.

To demonstrate the effect that the between score rank correlation has on aggregate totals, consider the following examples. In each case, the assignments are to be weighted equally. The raw score totals and the balanced score totals are calculated by summing the appropriate scores. For each

example, the critical value (alpha=.10; n=6) of the ranked correlation coefficient for raw score totals with balanced score totals is r_s= .829. The bracketted data are the balanced scores.

Student	Assignment 1	2	Total
A	100 [64.3]	40 [64.8]	140 [129.1]
B	83 [59.1]	31 [57.3]	114 [116.4]
C	58 [51.5]	26 [53.1]	84 [104.6]
D	31 [43.3]	15 [43.9]	46 [87.2]
E	25 [41.5]	12 [41.4]	37 [82.9]
F	21 [40.3]	10 [39.7]	31 [80.0]
X_m	53	22.3	
S.D.	32.9	11.9	
r_s (between) = +1			r_s (total) = +1

Table XXVIII - Effect of r_s for between scores on total score r_s when r_s (between) =+1

Table XXVIII displays the resulting rank correlation coefficient (r_s=+1) for raw score totals with balanced score totals when the between scores rank correlation coefficient is very large (r_s=+1). When the scores of the second assignment are rearranged to become rank ordered as 2,3,1,5,6,4 (as in Table XXIX), the between score rank correlation coefficient becomes +.657 while the rank correlation coefficient for aggregate raw scores with aggregate balanced scores becomes +.942.

Student	Assignment 1	2	Total
A	100 [64.3]	31 [57.3]	131 [121.6]
B	83 [59.1]	26 [53.1]	109 [112.2]
C	58 [51.5]	40 [64.8]	98 [116.3]
D	31 [43.3]	12 [41.4]	43 [84.7]
E	25 [41.5]	10 [39.7]	35 [81.2]
F	21 [40.3]	15 [43.9]	36 [84.2]
X_m	53	22.3	
S.D.	32.9	11.9	
r_s	.657		r_s= .942

Table XXIX - Effect of r_s for between scores on total score r_s when r_s (between) =.657

If the order of the second assignment is altered to 3,1,5,6,4,2 the r_s for between scores and the r_s for total scores become +.14 and +.83 repectively. Finally, if the ranks are changed to 6,5,4,3,2,1 then the r_s for between scores and r_s for total scores become -1 and +.31.

It is interesting to note that, with these examples, only the last has a rank correlation coefficient for raw score totals and balanced score totals that is not significant (ie. < .829) at the alpha=.10 level of significance (n=6). These examples show how changes in the rank correlation coefficient for between scores affect the rank correlation coefficient for raw score totals and balanced score totals. Considering a high majority of the sample in this study

compute totals by summing raw scores (see Table XVIII page 66) and considering class assignments are expected to correlate positively with each other, it becomes more apparent why the ranking of the raw score totals correlated highly with the ranking of the balanced score totals in this study, in spite of the many large differences in raw score variability.

Research Question 2

In response to the second research question regarding the reliability of the aggregation and weighting of raw scores for grading purposes, the sample's record sheets and the questionnaire responses pertaining to grading techniques were analyzed. The analysis revealed a high degree of similarity between the aggregation methods evident on the record sheets and the aggregation techniques described on the questionnaire.

Based on these results and on the rank correlation coefficients obtained (as discussed in Research Question 1), it can be concluded that this sample of subjects collated raw scores in a reasonably reliable manner.

The analysis of the data also indicated that most respondents did not attempt to apply weighting factors to raw scores. As indicated previously in this study (page 7), weighting raw scores and how much to weight is a matter for professional judgement. However, many subjects seemed oblivious to the fact that by not weighting, small and relatively insignificant assignments may easily contribute as

much or more to an aggregated score as does a major test or project. Most of the respondents in this study did not apply weights and thereby allowed each assignment to contribute its "natural weight" toward the aggregated score. Many of those who did attempt to weight the raw scores considered an assignment's total possible mark to be the weighting factor. For example, an assignment that has a total possible score of 100 was thought by many to contribute twice as much to an aggregate score as would an assignment that has a total possible score of 50. The example on page 9 of this study demonstrates the inappropriateness of this method as a means to weight raw scores. Assignments with a larger total possible value may possibly contribute less toward an aggregate score than does an assignment with a smaller total possible value. The key factors in reliable weighting are the standard deviation values rather than the total possible scores (see page 10). This notion was supported by such authors as Ahmann and Glock (see page 6), Ebel (see page 27), and Gronlund (see page 28). In light of these facts, the results of this study would suggest that the subjects' weighting techniques are less than reliable.

Unreliable weighting methods may have had a profound effect on the letter grades reported in this study. Since 95% of the subjects either relied on natural weightings or applied unreliable weighting methods (see Table XV), many students may have received an aggregate score that was not indicative of the emphasis placed on the assignments and course objectives.

For example, a minor assignment dealing with only a few course concepts could have had more influence on the aggregate score than did a comprehensive project that addressed several course objectives. However, if the raw scores had been properly balanced and weighted to reflect the emphasis of the course, then many students in this study would have likely received a different and more representative aggregate score. Furthermore, some students may have also received different letter grades based on the ranking of the new aggregate scores. Consequently, letter grades based on balanced and appropriately weighted raw scores have greater validity and reliability as indicators of the students' relative achievement.

Considering the possible consequences of unreliable grading practises (see page 2), the results of this study suggest that many students may have received misleading and inconsistent information in the form of unreliable letter grades. Decision makers, using these grades as a source of information, could be prone to making erroneous decisions. For example, the letter grades awarded some students may indicate that the course objectives have been successfully completed when in fact, the goals have not been met. In such instances, students could be promoted to the next course or level when such a promotion is not justified. The opposite could also occur. Students may be retained to repeat or review the course objectives when in fact the material has been well mastered and such a retention is not warranted.

Others may also be misguided by unreliable grades. For example, those who normally use student grades as a means of evaluating courses, programs, or pedagogical techniques, may implement unnecessary or incorrect modifications based on the unreliable information. In short, letter grades derived from unreliable weighting practises may be lacking in validity and reliability and may mislead others to formulate incorrect conclusions and decisions. Raw scores should be balanced and weighted in such a manner as to make the aggregate score representative of the emphasis placed on the course content and the various course objectives.

Based on the results of this study, it can be concluded that relatively few teachers weight raw scores reliably.

Research Question 3

In response to the third research question pertaining to the reliability and fairness of the sample's grades, the respondents' raw score grades and balanced score grades were compared. The results were recorded and tabulated in terms of the change that occurred to both letter grade and category when balanced grades replaced raw score grades. To assess fairness of the grades, the methods used to calculate aggregate scores for students who missed receiving a score on one or more assignments were analyzed.

These results would indicate that, in some

circumstances, students could be unfairly penalized or rewarded for missing an assignment. The following example demonstrates how a student might be penalized by missing an assignment when means and standard deviations are not considered.

	ASSIGNMENT SCORES (%)			
Student	1	2	3	Total
A	50	30	100	60
B	40	28	ab	34
C	30	26	74	43
D	10	20	60	30
Mean	33	26	78	
S.D.	17	4	20	

Table XXX - Compensation for missed assignment

Table XXX shows how student B has dropped to the third rank position on the basis of the first two assignment scores, even though he has consistently ranked second. By missing the third assignment with its higher mean score, student B has been penalized one rank position. On the other hand, if the missed assignment were calculated by averaging balanced scores (ie. T-scores), then his second place ranking would be preserved. The balanced score averages for students A, B, C, and D would be 60, 55, 49, and 37 respectively. Clearly, student B has not been penalized nor rewarded under the

balanced score method of compensation.

The results of this study would suggest that teachers' grades may lack reliability and, when students have missed assignments, the technique for compensation may not always be fair.

Research Question 4

In considering the fourth research question regarding the justification of the time and expense required to retrain teachers, the responses to the attitudinal section of the questionnaire were analyzed and the results tabulated in Tables XIII through XVII. These results, in conjunction with the previously discussed results, would indicate that in-service instruction and pre-service training in particular aspects of grading and reporting would be justified for many members of the research sample. Areas of greatest need appear to be those concerning the reliable weighting of raw scores, the reliable allocation of letter grades, and the reliable calculation of compensation scores for students who have missed assignments.

Current literature indicates that many teachers are uneasy or defensive concerning their grades and grading procedures (see page 5). During the course of this study, this notion was verified informally on several occasions. Comments, such as "...sure hope these marks are alright", or

"I've never been trained in this area", and other similar remarks were frequently made by the respondents. These comments and the uncertainty accompanying them would suggest that time devoted to pre-service and in-service would be well spent.

Weaknesses To Be Considered

The framework of the proposed study was considered to be sound. Many of the extraneous factors that could affect the results were controlled by randomization wherever possible. Although this project encountered few shortcomings, those that did exist were considered to be minor and not serious enough to adversely affect the results.

One of the most serious obstacles concerned the sample. Although of sufficient size for an adequate study, the sample was comprised entirely of volunteers. As a result, a bias may have been introduced into the study. The possible bias would result from the sample not being totally representative of the target population. Those who participated may have contributed a particular bias while those who declined to participate may have deprived the study of data unique to non-participants. For example, those who are particularly defensive about their grading procedures or about the reliability of their grades may not have volunteered. On the other hand, perhaps only those who are confident their grading

techniques are defensible or those who take a particular interest in marking and grading volunteered to become members of the sample. Teachers belonging to these and other such groups should be included in the sample since generalizations will be made about the population to which they belong. However, with volunteer samples, these variables are difficult to control.

Other weaknesses in this study concerned the questionnaire that was used to collect data on grading techniques. The question offering fictitious data to be collated and ranked posed particular problems for many of the respondents. Most recognized, from the information given in the directions, that some sort of weighting was required, even when they did not customarily weight raw scores. Some ignored the directions on weighting and applied their regular grading methods. Others attempted to accommodate the directions by developing new and unfamiliar strategies to weight the raw scores in the question.

The fictitious data section posed another weakness, although less serious and unrelated to the one previously discussed. An opportunity for the respondents to calculate an aggregate score for a student who had been absent should have been included in the fictitious data. This would have provided a third source of information for that particular aspect of grading. Although adequate, the record sheets and the written responses were the only sources of data for this topic. As a result of these weaknesses, the fictitious data

section in this study should be reviewed and revised before being used on other similar research projects.

Another minor weakness concerned the balanced raw score totals that were given to the respondents to regrade. Unlike raw scores, T-scores do not have a maximum possible score; rather, they are continuous. When the balanced T-scores were supplied to the respondents to be regraded, a maximum score should have been included for those who utilize per cent grading techniques. For those respondents who inquired, a score of 275 was provided as this would exceed all balanced score totals and therefore would yield percentages less than 100 per cent. Although this weakness caused some confusion among the respondents, it was not considered serious enough to flaw the results. Rather, it may have contributed toward a better understanding of the difficulty teachers experience when they assign letter grades to aggregated scores (see page 79).

Future Research

With the introduction of computers into most schools, the opportunities for improvement in the measurement, evaluation, grading, and reporting of student achievement have expanded dramatically and offer many new areas for future research. These areas have been divided into two main sections to facilitate discussion. The first area of research possibilities deals specifically with those topics related to

teachers' grading practises. The second area deals with measurement and evaluation of student achievement on a more global scale.

While this study touched only a few aspects of teacher grading practises, it left many questions unanswered. At the same time, it raised more questions to be resolved. Among the questions that still need to be addressed is the concern regarding the appropriateness of using T-scores in the classroom situation. The linear transformation of raw scores to T-scores first requires that the raw scores be converted to z scores. Further studies should be carried out to determine if the use of T-scores is appropriate for weighting and combining raw achievement scores.

Research opportunities exist in the replication of this study. One possibility would be to determine if location has an effect on teachers' grading practises. For example, a study could involve sampling teachers from different regions of the province to determine if differences in grading practises can be attributed to geographic location. It might be argued that teachers with superior grading practises are those who live in close proximity to universities.

Computerized record keeping and mark management is another area that should also be studied. If teachers are given computer programs that will reliably balance, aggregate, and weight raw achievement scores, would they be willing to alter their present grading practises for the newer computer approach? If it can be shown that teachers would readily

accept the new computer grading method, then studies would also have to be done to ascertain the most effective and useful computer program currently available.

This study has possibly identified a relationship between (a) the effects different standard deviations have on raw score totals and b) the rank correlation coefficients for between scores. The effects of large differences in variance appear to be diminished by large positive rank correlations for between scores. Studies should be done to determine if this relationship applies to all sets of raw scores and, if so, how the relationship could best be applied to classroom grading practises.

Another future research possibility relates to the rank correlation coefficients for between scores. Studies should be done to determine how well sets of classroom achievement scores actually correlate. As indicated in this study, it would be expected that the more capable student would normally attain the higher marks while the less able student would achieve the lower scores. The rank position of each student would therefore be relatively stable. Research should be conducted to determine if this is actually true for most classes.

The second area for future research, the general topic of measurement and evaluation of student achievement, also needs to be addressed. More studies need to be done to explore the reliability of teacher constructed evaluation instruments. Grading procedures can be "state of the art", but unless the

instruments used to measure student achievement are reliable and valid, student grades will still lack reliability, validity and fairness. Studies need to be undertaken to answer such questions as the following:

1. Are teacher constructed tests reliable and valid?
2. Are there ways to improve these instruments?
3. Are there better ways for teachers to construct measurement instruments?

These are questions that should be investigated.

Another possibility for future research involves the medium used to acquire measurement data on student achievement. With computers becoming more popular in schools, studies should be performed to determine if the reliability of classroom tests can be improved through the use of the computer.

With the availability of computers to assist teachers in evaluating student achievement, it is hoped that future research will offer new and more reliable methods of measuring, evaluating, grading, and reporting student achievement.

Bibliography

Ahmann, J. S., & Glock, M. D. (1981). Evaluating student progress - principles of tests and measurements (6th ed.). Boston: Allyn & Bacon.

Cohen, D. (1973). Evaluation reinterpreted. Australian Science Teachers Journal,19(2), 29-34.

Curwin, R. (1978). The grades of wrath: Some alternatives. Learning, 6, 60-64.

Ebel, R. L. (1972). Essentials of educational measurement. Englewood Cliffs: Prentice-Hall.

Ebel, R. L. (1974). Shall we get rid of grades?. NCME Measurement in Education, 5(4), 1-5.

Erling, A. S. (1979). Combining student test scores reliably. Educational Research and Methods, 11(2),42-50.

Gensley, J. T. (1969). A new method of evaluation for gifted students: A diary of learning. Gifted Children Quarterly, 13, 119-25.

Glass, G. V., & Stanley, J. C. (1970). Statistical methods in education and psychology. Englewood Cliffs, New Jersey: Prentice-Hall.

Gray, L. R. (1980). Educational evaluation and measurement - Competencies for analysis and application. Columbus: Merrill.

Green, H. A., Jorgensen, A. N., & Gerberich, J. R. (1962). Measurement and evaluation in the elementary school. New York: McKay.

Gronlund N.E. (1974). Improving Marking and Reporting in Classroom Instruction. New York: MacMillan.

Gronlund, N. E. (1981). Measurement and evaluation in teaching (4th ed.). New York: MacMillan.

Guilford, J. P., & Fruchter, B. (1973). Fundamental statistics in psychology and education (5th ed.). New York: McGraw-Hill.

Hills, J. R. (1981). Measurement and evaluation in the classroom. Columbus: Merrill.

Hopkins, K. D., & Stanley, J. C. (1981). Educational and psychological measurement and evaluation (6th ed.). Englewood Cliffs, New Jersey: Prentice-Hall.

Hopkins, C. D., & Antes, R. L. (1979). Classroom testing. Atasca, Illinois: Peacock.

Horrocks, J. E., & Schoonover, T. I. (1968). Measurement for teachers. Columbus:Merrill.

Isaacs, G., & Imrie, B. (1981). A case for professional judgment when combining marks. Assessment and Evaluation in Higher Education, 6, 3-25.

Johnson, D. M., & Van Osdol, B. M. (1974). A computer program for reliable grade assignment. Journal of Educational Data Processing, 11(2), 11-19.

Lien, A. J. (1976). Measurement and evaluation of learning. Dubuque, Iowa: Brown.

Lindvall, C. M. (1967). Measuring pupil achievement and aptitude. San Francisco: Harcourt, Brace & World

Lyman, H. B. (1978). Test scores and what they mean. Englewood Cliffs, New Jersey: Prentice-Hall.

Marshall, M. S. (1971). Why grades are argued. School and Society, 99, 350-353.

Mehrens, W. A., & Lehmann, I. J. (1984). Measurement and evaluation in education and psychology (3rd ed.). New York: CBS College.

Ministry of Education, Province of British Columbia. (1979). Construction and use of classroom tests. Victoria, BC: Author.

Ministry of Education, Province of British Columbia. (1979). Grading practises: Issues and alternatives. Victoria, BC: Author.

Nelson, C. H. (1970). Measurement and evaluation in the classroom. New York: MacMillan.

Ratzlaff, H. C. (1979). A to E approaches to grading student achievement. Vector, 21(1), 22-28.

Stewart, W. J. (1975). A multi-dimensional evaluating-reporting system in the elementary school. Reading Improvement, 12, 174-76.

Tenbrink, T. D. (1974). Evaluation - A practical guide for teachers. New York: McGraw-Hill.

Townsend, E. A., & Burke, P. J. (1975). Using statistics in classroom instruction. New York: MacMillan.

Wick, J. W. (1973). Educational measurement - Where are we going and how will we know when we get there. Columbus: Merrill.

Appendix A

The following questionnaire was used to collect information on the subjects' demographic variables, on their attitudes and perspectives toward grading, and on their grading practises and techniques. It was distributed to each subject at the end of the data collection period.

Respondent # _____

GRADING TECHNIQUES SURVEY

Section I - Demographic Information

DIRECTIONS: Circle the letter that corresponds to the correct or most appropriate response.

1. How many years teaching experience do you have? (combine both public and private school experience)
   a. 0-5
   b. 6-10
   c. 11-15
   d. 16-20
   e. more than 20

2. What is your teaching assignment?
   a. full time
   b. part time

   If "part time", describe your teaching assignment (percentage, subjects taught, etc.)

   _____

   _____

3. How many Statistics courses or Measurement and Evaluation courses have you completed? (disregard unit value)
   a. none
   b. one
   c. two
   d. three
   e. more than three

4. How many years of university have you completed?
   a. two
   b. three
   c. four
   d. completed Bachelor of Education degree
   e. completed MA or MEd degree
   f. other
   (explain:_____)

5. Which is your age category?
   a. less than 25 years
   b. 25-30 years
   c. 31-40 years
   d. 41-50 years
   e. 51-60 years
   f. greater than 60 years

Section II - Grading and Reporting

DIRECTIONS: You are to express, on a five-point scale, the extent of agreement between the feeling expressed in each statement and your own personal feeling. The five points are: Strongly Disagree (SD), Disagree (D), Undecided (U), Agree (A), Strongly Agree (SA). You are to encircle the letter(s) which best indicates how closely you agree or disagree with the feeling expressed in each statement.

1. My colleagues at this school take report card marks very seriously.

   SD D U A SA

2. Letter grades are an effective method of informing parents of their child's progress and achievement.

   SD D U A SA

3. To the best of my knowledge, parents are generally satisfied with the letter grade system of reporting.

   SD D U A SA

4. I am very confident that the letter grades I assign are accurate and reliable.

   SD D U A SA

5. I would like to learn more about collating and assigning letter grades to raw scores.

   SD D U A SA

6. More in-service sessions on grading and reporting should be offered.

   SD D U A SA

7. Universities should offer more pre-service instruction in effective grading and reporting.

   SD D U A SA

8. The report card format allows sufficient information to be communicated to parents, students, etc.

   SD D U A SA

DIRECTIONS: Answer the following in the space provided.

10. Briefly describe how you combine assignment and test scores to determine letter grades. (Use back of page if necessary)

_____

_____

_____

_____

_____

11. In reference to the method described in #10, which subjects are graded this way? (all subjects, mathematics, language arts, science, social studies, art, P.E., music, French)

_____

12. Briefly describe how you arrive at a final term score for a student who has been absent for one or more tests or assignments. In other words, how do you compensate for a legitimately missed assignment? (Use back of page if necessary)

_____

_____

_____

_____

_____

DIRECTIONS: The following record sheet represents the raw scores (percent) that five students obtained on five assignments. Each assignment had a total possible score of 100. The first assignment represented about 10% of the course; the second and third about 20% each; the fourth about 10%; and the fifth, the final test, about 40%. Use the methods you normally use (as indicated in PART II #10 above) to calculate and record the Total Score for each student. Indicate the rank of each student (1 = highest; 2 = next highest; etc.) in the space provided.

	ASSIGNMENTS raw scores (%)						
NAME	1	2	3	4	5	SCORE	RANK
1. Student A	50	55	65	57	95		
2. Student B	69	60	58	80	83	_____	_____
3. Student C	45	72	52	71	75	_____	_____
4. Student D	75	95	45	83	25	_____	_____
5. Student E	73	83	60	63	36	_____	_____

Thank you for taking the time to complete this survey.