An NCME Instructional Module on

Developing a Personal Grading Plan

David A. Frisbie and Kristie K. Waltman
University of Iowa

The purpose of this instructional module is to assist teachers in developing defensible grading practices that effectively and fairly communicate students' achievement status to their parents. In formulating such practices, it is essential that teachers first consider their personal grading philosophy and then create a compatible personal grading plan. The module delineates key philosophical issues that should be addressed and then outlines the procedural steps essential to establishing a grading plan. Finally, the features of several common methods of absolute and relative grading are compared.

This instructional module has been designed to help prospective and beginning teachers sort out the issues involved in formulating their grading procedures and to help experienced teachers reexamine the fairness and defensibility of their current grading practices. It can be applied at any grade level and in any subject matter area in which letter grades are assigned to students at the end of a reporting period. The content focus is limited to grading, so other modes of evaluating and reporting student progress are not addressed.

With regard to the purpose of grades, the position we will assume and defend is that grades are intended mainly to communicate the achievement status of students to their parents. The grade, then, symbolizes the extent to which a student has attained the important instructional goals of the reporting period for which the grade is assigned. Grades would not be needed if there were no need to communicate achievement to parents (or others outside the school setting). Grades are not essential to the instructional process: teachers can teach without them and students can and do learn without them.

Grades do serve several other important functions that are secondary to their school-to-home communication role, however. Grades provide incentives to learn for many students. Most students are motivated to attain the highest grades and to receive the recognition that often accompanies such grades, and they are motivated to avoid the lowest grades and the negative outcomes that sometimes are associated with those grades. Grades also provide information to students for self-evaluation, for analysis of strengths and weaknesses, and for creating a general impression of academic promise, all of which may enter into educational and career planning. Finally, grades are used to communicate students' performance levels to others who want to know about past achievement or want to forecast future academic success. Prospective employers and teachers in subsequent classes use grades in these ways. So do those who are charged with deciding who qualifies for honor society, who is eligible for basketball, or who should be the class valedictorian.

This module is organized to demonstrate the process a teacher might follow in devising a grading plan. First, some of the philosophical issues inherent in the grading process are identified, and then steps to follow in creating a grading plan are outlined. Finally, some of the most common methods of assigning grades are analyzed. The primary objectives of this module are to enable the reader to (a) describe the main questions of value that need to be considered in formulating a personal grading philosophy; (b) explain how written district grading policies, district reporting forms, and building-level expectations can help or hinder the development of a personal grading philosophy; (c) identify the essential procedural questions that need to be resolved in developing a personal grading plan; (d) explain how the decisions about defining the grade symbols directly influence other subsequent decisions in creating a personal grading plan; and (e) analyze the strengths and weaknesses of each of several common methods of assigning grades.

Teachers who implement the recommendations of this module should end up with a defensible grading plan that is in harmony with their personal grading philosophy and the grading policy of the district in which the plan will be implemented.

David A. Frisbie is a professor of measurement and statistics in the College of Education at the University of Iowa, 316 Lindquist Center, Iowa City, IA 52242.

Kristie K. Waltman is a doctoral student in the College of Education at the University of Iowa.

Series Information

ITEMS is a series of units designed to facilitate instruction in educational measurement. These units are published by the National Council on Measurement in Education.

Developing a Grading Philosophy

The process of grading requires teachers to make a number of decisions that are grounded in their personal value system. What to do about grading or how to do it is often less a matter of correctness and more a matter of preference and perceived value or importance. In this section, we identify a number of

"should" questions, questions about which reasonable people might disagree because of their personal beliefs, values, and experiences. "What should a B mean? Should any student be assigned an F grade? How many A grades should be assigned in a class?" These are questions for which research studies cannot provide answers, but they are the types of questions that must be answered by each teacher who issues grades.

1. What meaning should each grade symbol carry? A grade of C can tell how much Rudy knows, how he compares to his classmates, how hard he has tried, how much he has learned this quarter, or how well he has behaved this term. Since it cannot tell all of these things at once, what should it be limited to telling?

2. What should "failure" mean? There is undoubtedly more emotion associated with the F grade than any other, largely because of the negative consequences for many students who receive it. What does F mean? Should it mean the student knows nothing, knows the least within his class group, can do only the lowest level of work in the curriculum, hasn't tried to learn, or hasn't learned much in 9 weeks?

3. What elements of performance should be incorporated in a grade? Once a teacher has decided on the meaning the grade symbols should convey, much effort will be required to keep contaminating information out of the grade. Teachers are constantly making observations and judgments about a variety of characteristics of their students. Such information can be used to evaluate communication skills, interpersonal relations, attitude, and motivation, but not all information gathered need be funneled into the grading decisions. What should be included and what should be kept out?

4. How should the grades in a class be distributed? In some districts, written grading policies dictate the nature of grade distributions (e.g., the percentages of As, Bs, etc.); however, most districts seem not to have such policies. Thus, most teachers are probably faced with a decision about the percentage of A grades or C grades they should issue. Should the average grade be C? Is it okay if everyone gets an A? Should there be an equal number of B and D grades?

5. What should the components be like that go into a final grade? The separate scores or grades that are combined to form the final grade for a reporting period must, above all, convey the meaning the teacher previously decided upon for the grade symbols. Should rough drafts count? How about scores from a test that turned out to be too hard? What about practice trials for performance tests? How many components should there be as a minimum?

6. How should the components of the grade be combined? Suppose Mr. Voss uses three tests, a short paper, and an individual project for third quarter grading in his sixth-grade social studies class. Should each of the five components be worth 20% of the final grade or should some be more heavily weighted? What should he think about when making that decision?

7. What method should be used to assign grades? After component scores have been combined, a final grade needs to be assigned to each student. The method of assignment ought to be consistent with the decisions made earlier about the meaning each grade symbol should have. For example, it would be illogical to grade on the curve if grades are to be based on absolute standards of performance. Which of the several methods of absolute grading is best?

8. Should borderline cases be reviewed? If borderline cases are to be reexamined to decide on the appropriateness of the grades, here are some questions the teacher needs to address: How close to a cutoff point does a score need to be before it is considered borderline? Should only grades just below a cutoff be checked or should those just above be looked at also? What additional information should be examined to help make the borderline decisions? Should students be allowed to furnish extra credit work to raise a borderline grade?

9. What other factors can influence the philosophy of grading? A teacher's personal philosophy of grading also can be shaped by school district grading policies and building practices. For example, some district grade-report forms provide descriptive phrases to define each grade symbol. In such cases, written district policy is inherent in the reporting form even though grading procedures are not prescribed explicitly. In the absence of written policy, however, the most recent grades issued become the norm; practices that depart noticeably from the norm are likely to be squelched, regardless of the philosophy of the grader.

Establishing a Grading Plan

This section of the module details the sequential steps involved in applying a personal philosophy of grading to form a personal grading plan. It is a personal plan because it incorporates the personal values, beliefs, and attitudes of the particular teacher who will use it to assign grades. And though a philosophy of grading is the foundation for establishing a grading plan, the plan is also shaped and influenced by current research evidence, prevailing lore, reasoned judgment, and matters of practicality.

Step 1. Identify and implement written district policy. If there is written district policy on grading, teachers are obligated professionally (and probably legally) to follow it. The policy may be in the form of detailed rules or it may be a set of general statements from a school board resolution. It may simply be reflected in the reporting form sent to parents, in the statements of purpose on the report card, or in the explanations of the meanings of the grade symbols used.

What should you do if your philosophy and preferred grading procedures conflict with written policy? First, a discussion with your building administrator may be the most reasonable approach because the administrator is the first line of enforcement of district policy. If the results of such a meeting are not satisfactory, a next step would be to follow the existing policy while informally surveying your colleagues to see whether they would support a change. If so, efforts to alter the policy to fit the philosophies of the staff could be very productive.

Step 2. Decide what the meaning of each grade symbol will be. There are three facets to the meaning of a letter grade, and the teacher needs to make a decision about each facet for his or her plan. First, the grade compares performance either to a relative standard (norm-referenced) or to an absolute standard (criterion-referenced). For example, a relative comparison is being made if a C grade means "average performance compared to others in the class," but an absolute comparison is being made if it means "demonstrated attainment of the most important objectives." It is essential for the teacher who adopts a criterion-referenced meaning to develop a description of the student behavior that defines each grade symbol. Figure 1 illustrates the types of phrases that can be used to differentiate levels of performance on the absolute grading scale. These phrases are contrasted with descriptors of relative grades that depend entirely on average performance to obtain their meanings. Note that to describe a "B student" using absolute standards, no reference is made to the achievements of other students. Instead, the comparison is based on the knowledge and skills studied and the extent to which prerequisites for future learning have been attained. The selection of a relative or an absolute grading standard is very critical because, once that selection is made, all of the tools of assessment that are used to obtain grading information should be designed in accord with that selection—either norm-referenced or criterion-referenced.

A second facet of the meaning of a grade indicates whether achievement or effort is being described. Obviously effort and

Grade	Absolute Scale, Criterion-referenced	Relative Scale, Norm-referenced
A	• Firm command of knowledge domain • High level of skill development • Exceptional preparation for later learning	Far above class average
B	• Command of knowledge beyond the minimum • Advanced development of most skills • Has prerequisites for later learning	Above class average
C	• Command of only the basic concepts of knowledge • Demonstrated ability to use basic skills • Lacks a few prerequisites for later learning	At the class average
D	• Lacks knowledge of some fundamental ideas • Some important skills not attained • Deficient in many of the prerequisites for later learning	Below class average
F	• Most of the basic concepts and principles not learned • Most essential skills cannot be demonstrated • Lacks most prerequisites needed for later learning	Far below class average

FIGURE 1. *Descriptors of grade-level performances using absolute or relative standards*

achievement are not independent, but a single grade cannot describe both unambiguously. Ideally, separate grades or marks should be used for each trait so that the two can be described more purely at the same time. If only one grade can be issued, however, describing achievement rather than effort seems more beneficial.

The third facet is a time-related reference—growth vs. status. If a grade is to indicate the amount of growth from the beginning of the grading period until the end, the highest grades should be assigned to those who demonstrate the greatest gains. In many subject areas, those with high beginning achievement levels will likely be able to grow the least. In fact, in some units of instruction, the highest achieving student may grow very little, if at all. But, assigning a C or D grade to such a student seems counter to the general notion of what grades usually connote. In short, most parents, students, and teachers are interested in whether growth has occurred, as they should be. But more important to them is the level of achievement at a particular time and whether that level is sufficient for moving onto the next sequence of the instructional program.

Step 3. Check the grade meanings against your instructional approach for logical consistency. A teacher who uses an outcomes-based approach or a highly individualized approach to instruction would not logically choose to use grades that have a norm-referenced meaning. Another teacher who depends heavily on the principles of cooperative learning would not likely use norm-referenced grades because of the competition they breed. Teachers who are devoted to a specific instructional or teaching philosophy need to develop a grading plan that is compatible with their teaching philosophy.

Step 4. Identify evaluation variables, reporting variables, and grading variables separately. The interpretability of a course grade will be jeopardized if the grade is made to carry too many pieces of information. This is the main reason why effort should be separated from achievement and growth should be separated from status when establishing the meaning of each grade symbol. Failure to make these separations will introduce irrelevant noise; static in any communication leads to misunderstanding and subsequent inappropriate decisions and impressions. One way of guarding against the threats to clear communication involves planning for evaluation. That is, just as plans should be made about what to teach and how to teach, concurrent plans should be made about the type of evaluation information that should be gathered during instruction.

Teachers gather preinstructional information about students' entering behaviors, they gather additional information to monitor student and class progress, and they obtain further information to decide if students are ready to move on to a new instructional unit. Thus, the *evaluation variables* that teachers depend on include such learner characteristics as interests, preferences, academic ability, past achievements, attitudes, effort, conduct, study skills, interpersonal skills, and the like. There are too many such variables to enumerate, but teachers can identify many of them and make definite plans to gather information about them. But having gathered such a wealth of information, it is not their intention to report the outcomes or judgments about all of them to parents or students. Ordinarily they select a small subset of such variables, which can be called *reporting variables,* as required by the district reporting methods, and they will use symbols or narrative comments to pass on the selected information.

Finally, from the set of reporting variables described above, a teacher will select those that provide information that is consistent with the meaning of the grades the teacher plans to assign. This subset of reporting variables can be labeled *grading variables.* The teacher who is determined to use grades to describe achievement levels will temporarily set aside indicators of effort, demeanor, attitude, and congeniality in favor of performance assessments and scores on tests, papers, and projects. The latter reflect achievement more accurately.

Note that it is possible to distort the meaning and value of certain grading components that, on the surface, appear to be relevant grading variables. For example, if the social studies essay scores of some students are reduced because of deficien-

Table 1
The Distribution of Instructional Objectives Within Grading Components for Three Units of Instruction

	Unit 1		Unit 2		Unit 3	
Grading Component	Objs. 1–12	(33%)	Objs. 13–24	(33%)	Objs. 25–36	(33%)
Tests	1–8	(22%)	13–20	(22%)	25–32	(22%)
Quizzes	3–5		13–15		25–27	
Lab reports	9–10	(5%)	22–24	(6%)	36	(3%)
Homework	8–9		23–24		35–36	(2%)
Lab practicals	11–12	(5%)	23	(3%)	34	(3%)
Performance tests	—		21	(3%)	33,36	(2%)

cies in writing mechanics, how well do those scores describe achievement in social studies? If the teacher assigns an A to a group project, what does that A mean for a member of the group who made little contribution to planning, conducting, or summarizing project activities? If the grade on a paper is dropped a full letter for each day it is late, what does the final grade on a late paper indicate about achievement in language arts? If a student has an unexcused absence on the day of a test, what does an F grade for that test contribute to a quarter grade that is supposed to describe achievement? This is not the place to argue the merits of such policies or to explore alternative actions, but it is germane to point out that "relevant" grading variables can be distorted. Tainted component scores cause tainted composites. Tainted composites lead to misinterpretation.

Step 5. Check to see what the grade distributions in your building have been like at your grade level in the subjects you teach. If no written district policy exists, the grades issued in the most recent years will be the norm against which the reasonableness of each teacher's grades will be judged. How would your principal (and other teachers) react if your outcomes-based approach resulted in A grades for all of your students? This hypothetical question can not be answered, but it points out that grading patterns that depart significantly from local history generally will be questioned.

Suppose you teach an honors class in algebra and also have a regular algebra class. Should the grade distributions be similar in the two classes? If the grades from the two classes were merged into a single distribution, should that large distribution have the same number of A grades as would be assigned in two regular classes (assuming no honors section)? If written policy does not speak to these issues, the grades from the past few years are probably the best indication of what the current outcomes should be like.

Step 6. Decide on the kinds and number of grading components needed. Is it reasonable to base a 9-week English grade only on the score from a single test? Most would say, "Definitely not." Would scores from only two tests be sufficient? "Better," most would probably say, "but far from ideal." Generally, the more *good* information available for assigning grades, the more likely those grades will represent actual achievement levels accurately. There is no minimum number of tests or other grading components that should be used; the overriding concern is to assess attainment of as many of the instructional objectives as possible so that grades will represent accomplishments with respect to the entire domain. The types of grading components required should be determined by examining what the instructional objectives require.

At this stage, it is also important to rule out the use of certain achievement-oriented evaluation variables from the set of grading variables. All of the instructional activities and exercises that students complete for practice purposes should be regarded as evaluation variables that inform teachers about progress *during* learning, not status indicators at the *end* of a learning experience. Daily homework, periodic quizzes, and responses to oral questioning are examples of evaluation variables that generally should not be regarded as grading variables. As long as a grade is intended to describe achievement status at the end of an instructional segment, assessments designed mainly to monitor progress during instruction should be excluded.

Should the contribution of individual students to a group project be factored into the grading of the project or the quarter's work? Can individual contributions be teased out? Should all group members be assigned the same grade? Should teachers simply provide evaluative reactions to group work but not treat such results as grading variables? Surely a student's grade should not be embellished or tainted by the achievements of others. Again, tainted composites lead to misinterpretation.

Many assessment techniques require a particular communication skill—writing, reading, speaking, drawing—that may not be well developed in some students. For example, a preponderance of essay testing may favor good writers, or the use of only objective tests may disadvantage poor readers. Obviously, students with limited English proficiency will be at a disadvantage no matter which medium of communication is used. The components of a grade ought to be selected or developed so that achievement in the subject area of interest (e.g., social studies) will not be masked by the language skills required by the assessment method.

Step 7. Determine how much weight each grading component will have. The role of instructional objectives is central to the process of combining grading components, just as it is for deciding which components to use. The task of formulating weights involves deciding how important each component score or grade is in describing achievement at the end of a grading period. The information in Table 1 illustrates the process of determining weights.

Table 1 shows that three science units were completed during one quarter, each unit consisted of 12 objectives, and each unit was to have equal weight (about 33%) in the quarter grade. The objectives measured by each grading component are identified by their number. Here is the initial thinking for determining the weights for the components of Unit 1:

1. Since 8 of the 12 objectives were covered by the test, two-thirds (8/12) of the weight for Unit 1 (33%) should be designated for the tests (22%).
2. The objectives measured by the quizzes were also covered by the test. Since they were regarded as checks during the learning process, the quizzes should have zero weight.

3. The lab reports covered 2 of the 12 objectives (17%), and no other grading component measured those same objectives. Give 17% of the unit weight (33%) to lab reports (about 5%).
4. Homework, like quizzes, was considered practice and dealt with objectives measured by other components. The quality of homework could be influenced by the help of others or it might be copied. No weight should be given to homework in the final grade.
5. Lab practicals, like lab reports, covered two unique objectives. Therefore, the same rationale was used to allocate 5% weight to lab practicals.

What factors entered into the thinking about component weights in the scenario above? One factor was the importance of the component as indicated in part by the number of objectives it encompassed. Another factor was uniqueness. Two components that measured any objectives in common were given less weight individually than two components that measured an equal number of unique objectives. (Notice how Objective 36 in Table 1 was handled.) A third factor, not evident in the scenario or Table 1, is the accuracy of the scores obtained from a component. For measures of similar skills, the one that provides the most accurate scores ought to be given the most weight.

Step 8. Determine how components will be combined to create a composite score or final grade. Once component weights have been established, the teacher must decide how to combine components so that the desired weights and actual weights are the same. The considerations and procedures for proper weighting differ for the norm-referenced and criterion-referenced situations. The differences are detailed in another instructional module and will not be repeated here (Oosterhof, 1987). For norm-referenced purposes, the variability of the scores of each component influences the weight the component will have in the composite. For criterion-referenced purposes, it is the total points associated with each component that matters most.

Step 9. Choose a method for assigning grades. The relative merits of the various common methods of assigning grades to composite scores are reviewed below. At this stage of establishing a grading plan, it is important for the teacher to choose or adapt a method of grade assignment that is consistent with the meaning that the grade symbols are intended to carry. Unfortunately, some of the most common methods of assigning grades yield results that are neither norm-referenced nor criterion-referenced. Consequently, teachers need to look carefully at methods of grade assignment that seem worthy of adoption.

The final aspect of assigning grades is the matter of dealing with borderline grades. For some teachers, the question is not *how* to treat borderline cases; it's *whether* to do it at all. They regard their grading practices as rigid procedures that produce highly objective grade results. For them, a review of borderline cases could insert subjectivity into the process and lead to outcomes that they would feel uncomfortable defending. However, others are driven by the apparent subjectivity inherent in several aspects of the grading process and by the desire to be fair in grading. Their notion of fairness is to err in favor of the student (award the higher of two grades) if an error is going to be made. The reconsideration of borderline cases, then, is one way to ensure that certain errors will not be too influential in determining a student's grade.

What basis should be used for deciding whether to raise a grade in a borderline situation? Nearly always, *achievement* information that was not used to assign the tentative final grade should be taken into consideration. This advice is consistent with the premise that a grade should describe achievement rather than effort or some other trait. Homework quality, quiz score average, quality of class participation, and contributions to cooperative learning experiences are all possible achievement-oriented evaluation variables that could be suitable for borderline reviews. Some teachers hold one high-quality piece of achievement data in reserve for just such purposes.

Some Relative Grading Methods

Grades derived from any of the relative grading methods will have certain shortcomings that are inherent in any grades intended to have a norm-referenced meaning. For example, unless the person interpreting the grade knows which reference group was used, the grade means very little. Was it the student's class, a combination of classes, or classes from the past two years? Further, by definition, a norm-referenced grade does not tell what a student can do; there is no content basis other than the name of the subject area associated with the grade.

Grading on the Curve

The curve referred to in the name of this method is the normal, bell-shaped curve that is often used to describe the achievements of individuals in a large heterogeneous group. The idea behind this method is that the grades in a class should follow a normal distribution, or one nearly like it. Under this assumption, the teacher determines the percentage of students who should be assigned each grade symbol so that the distribution is normal in appearance. For example, the teacher may decide that the percentages of A through F grades in the class should be 10%, 20%, 40%, 20%, and 10%, respectively.

Since some teachers who use the method rightly believe that classroom groups are too small for their achievement scores to resemble a normal curve, they choose percentages that, in their judgment, are more realistic. So they may decide on 20%, 35%, 30%, 10%, and 5%. The percentages are selected arbitrarily and are treated like grade quotas so that the top 20% of students in terms of their composite scores will earn an A, the next 35% would be assigned a B, and so on.

Grading on the curve is a simple method to use, but it has serious drawbacks. The fixed percentages are nearly always determined arbitrarily, and the percentages do not account for the possibility that some classes are superior and others are inferior relative to the phantom "typical" group the percentages are intended to represent. In addition, the use of the normal curve to model achievement in a single classroom is generally inappropriate, except in large required courses at the high school and college levels.

Distribution Gap Method

When the composite scores of a class are ranked from high to low, there will usually be several short intervals in the score range where no student actually scored. These are gaps. This method of grade assignment involves finding the gaps in the distribution and drawing grade cutoffs at those places. For example, if the highest composite scores in a class were 211, 209, 209, 205, 197, 196, . . ., then the teacher might use the gap between 205 and 197 to separate the A and B grades. The gap between 211 and 209 is too small and might produce too few A grades. The one between 209 and 205 might be large enough, but 205 seems more like 209 than 197.

In some score distributions there are many wide gaps; in others there are only a few narrow gaps. The sizes and locations of the gaps are determined by random errors of measurement as well as by actual differences among students in achievement. For example, Mike's 197 maybe would have been 203 (if there had been less error in his scores), and Theo's 205 maybe would have been 200. Under those circumstances, the A–B gap would be less obvious, and too many final grade decisions would have been made by reviewing borderline cases.

When gaps are wide enough, this method helps the teacher avoid disputes with students about near misses. But when the gaps are narrow, too much emphasis is placed on the borderline

information, information that the teacher had decided was not relevant enough or accurate enough to be included among the set of grading components that formed the composite. Only occasionally will the gap distribution method yield results that are comparable to those obtained with more dependable and defensible methods.

Standard Deviation Method

This relative method is the most complicated computationally, but it also is the fairest in producing grades objectively. It uses the standard deviation, a statistic that tells the average number of points by which the scores of students differ from their class average. It is a number that describes the dispersion, variability, or spread of scores around the average score. In this method, the standard deviation is used like a ruler to identify grade cutoff points.

Suppose you have formed composite scores for your class of 25 students and that the average was 129 and the standard deviation was 10. (Consult an introductory measurement or statistics book to see how to compute these statistics simply.) Assuming C to be the average grade, we can find the cutoff between B and C by adding, for example, one-half of the standard deviation to the average (129 + (0.5)(10) = 134). Then the A–B cutoff is found by adding 1.5 standard deviations (for example) to the average (129 + (1.5)(10) = 144). By subtracting corresponding values from the average score, the C–D cutoff is found to be 124, and the D–F cutoff is 114. (Can you verify these values?) The ranges for each grade are the following: A = 145 and up, B = 135–144, C = 124–134, D = 123–114, and F = 113 and below. These ranges can be made smaller or larger for groups of higher or lower ability level by adjusting the number of standard deviations used to find the cutoffs. For a particularly able class, for example, the A–B cutoff might be only one standard deviation above the average and the B–C cutoff might be 0.3 above, rather than 0.5.

Unlike grading on the curve, this method requires no fixed percentages in advance, and unlike the distribution gap method, the cutoff points are not tied to random error. When the teacher has some notion of what the grade distribution should be like, some trial and error might be needed to decide how many standard deviations each grade cutoff should be from the composite average. When a relative grading method is desired, the standard deviation method is most attractive, despite its computational requirements.

Some Absolute Grading Methods

Absolute grading methods produce grades that share some general shortcomings, independent of the particular method that generated the grades. For example, unless they are accompanied by a description of the performance standards or the content domains that have been studied, the meaning of an absolute grade is obscure. Furthermore, no criterion-referenced grading method produces grades that are strictly absolute in meaning. Such grades are based on performance standards that nearly always have a normative basis. A "B writer" in fourth grade should be able to use quotations in dialogues, the teacher may say, but if most fourth-grade students do not and cannot, the standard is likely to be lowered to reflect reality (the norm). Note that adjusting grades instead of modifying the standards would contribute to meaningless grades.

Fixed Percent Scale

This method uses fixed ranges of percent-correct scores as the basis for assigning grades to the components of a final grade. A popular grading scale is the following: 93–100 = A, 85–92 = B, 78–84 = C, etc. These ranges are fixed at the beginning of the reporting period and are applied to the scores from each grading component—written tests, demonstrations, papers, and performance assessments. Component grades are then weighted and averaged to get the final grade.

Unfortunately, a percent score will be meaningless unless the domain of tasks, behaviors, or knowledge upon which the assessment was based is defined explicitly. That is, a test score of 100% should mean that the student has complete or thorough attainment of the key elements of the area of knowledge that was sampled by the test. But if an assessment is developed in such a way that the underlying content domain is ill-defined or nebulous, the percent-correct scores from it will have no meaning beyond the specific tasks that comprise the assessment. Scores of 80% on a spelling test and 75% on a speech say little about performance unless we know the difficulty of the domain of spelling words and which important criteria were used to score the speech. In sum, percent scores cannot provide a reference to absolute performance standards unless the underlying knowledge domain is adequately described.

Another serious drawback of this grading method is the fact that the percent-score ranges for each grade symbol are fixed for all grading components. For example, the fact that 93% is needed for an A places severe and unnecessary restrictions on the teacher when he or she is developing each assessment tool. If the teacher believes there should be some A grades, a 20-point test must be easy enough so that some students will score 19 or higher; otherwise there will be no A grades. This circumstance creates two major problems for the teacher as assessment developer. First, it requires that assessment tasks be chosen more for their anticipated easiness than for their content representativeness. As a result, there may be an overrepresentation of easy concepts and ideas, an overemphasis on facts and knowledge, and an underrepresentation of tasks that require higher order thinking skills. The teacher may need to "fudge" on the domain definition to accommodate the fixed grading scale.

A further limitation of this method relates to the accuracy of the assessment information obtained. Since the grade cutoff scores usually are located between the 60% and 100% points on the percent scale, most of the scale points (0–60) are of no value in describing the different absolute levels of achievement. For example, if A and B performance must be in the range of 85–100%, the very best B achievement and the very worst B achievement are separated by only eight points (85–92), as are the very best and very worst A achievements (93–100). These are fairly narrow score ranges, especially considering the fact that a 100-point scale is available for use. Because these ranges are narrow *and* fixed, they will contribute to fairly inaccurate grades when the scores of any single grading component are not very dependable. If the grade ranges could be made larger when the scores of a certain component are fairly inaccurate, then more accurate grades would probably result.

The fixed percent scale method usually produces grades that have little meaning in terms of content standards, and it often yields grades that are of questionable accuracy. The percent cutoffs for each grade are arbitrary and, thus, not defensible. Why should the cutoff for an A be 93, 92, or 90? Further, why shouldn't the A cutoff be 88% for a certain test, 91% for another, and 83% for a certain simulation exercise? Is there any reason why the same numerical standards must be applied to every grading component when those standards are arbitrary and void of absolute meaning?

Total Point Method

Some teachers accumulate points earned by students throughout a reporting period and then assign grades to the point total at the end of the period. First the teacher decides which components will figure into the final grade and what the maximum point value of each component will be. (This is done before tests are developed and before the scoring criteria for projects are established.) For example, you may decide to use

two tests (50 points each), two papers (40 points each), and a report (20 points) for a maximum of 200 points for the quarter. Then the grade cutoffs might be set as follows: 180–200 = A, 160–179 = B, 140–159 = C, 120–139 = D, and 0–119 = F. Implicit in this set of ranges is a percent scale with grade cutoffs of 90%, 80%, 70%, and 60%. All teachers who use this method do not necessarily adopt these same cutoffs, but it is easy to see that there is no rational way to set the cutoffs. They are as arbitrary, and nearly as meaningless, as those derived from the fixed percent scale method. Unlike the fixed percent scale method, however, grades are not assigned to components with the total point method. And unlike grading on the curve, the arbitrary cutoff points are established at the beginning of the reporting period, *before* assessment results are known.

One of the difficulties of using this method is that often a decision has to be made about the maximum score on a project or test before the teacher has had ample time to think about the key ingredients of the assessment. Here's how this circumstance can contribute to poor assessment development practices: Suppose I need a 50-point test to fit my grading scheme, but I find as I build the test that I need 32 multiple-choice items to sample the content domain thoroughly. I find this unsatisfactory (or inconvenient) because 32 does not divide into 50 very nicely (It's 1.56!). To make life simpler, I could drop 7 items and use a 25-item test with 2 points per item. If I did that, my point totals would be in fine shape, but my test would be an incomplete measure of the important unit objectives. The fact that I had to commit to 50 points prematurely dealt a serious blow to obtaining meaningful assessment results.

Another potential drawback to the total point method is the ease with which extra credit points can be incorporated to beef up low point totals. This practice can simultaneously distort the meaning of the content domain and final grade. When the extra tasks are challenging and relevant to current instruction, this seems like a reasonable way to individualize and motivate high achieving students. In such cases, the outcome is likely to make high point totals even higher. But extra credit that simply allows students to compensate for low test scores or inadequate papers is not reasonable, especially if the extra work does not help them overcome demonstrated deficiencies. The point here is that this method of grading makes it convenient for teachers to allow extra credit work of the latter form to compensate for low achievement. When that happens, the grades take on a new meaning because the relevant domain of knowledge and skills gets redefined by the nature of the extra credit tasks.

Content-Based Method

This method involves assigning a grade to each component of the final grade and then weighting the separate grades to obtain the final one. The teacher develops brief descriptions of the achievement levels (standards) associated with each grading symbol, somewhat like those shown in Figure 1. These standards for "A work" and "B work" and so on are then used to establish the grade cutoff scores for every component. Compared to the fixed percent scale method, which keeps cutoff scores constant for all components, this method keeps the performance standards for a grade constant but lets the cutoff scores change. Here is an example of how the method might be used:

Suppose you have prepared a 30-item test to measure the achievement of most of the objectives in a unit of instruction. Assuming that grades A through F will be assigned to test scores, you will need to develop a brief description of the performance levels you expect students to reach for each of the five possible grades. For example, you might describe C expectations as "knows basic concepts and can do the most important skills; lacks some prerequisites for later learning." Using descriptions like these, you can begin an item-by-item review of the test.

For question no. 1, ask whether a student with only minimum achievement (D) should be able to answer correctly. If so, record a D next to the item; if not, pose the same question for grade C achievement. This process continues until the first item has been classified. For items that the teacher believes most A students will not necessarily answer correctly, a symbol such as N can be used to indicate that no grade level applies. (For items worth more than a single point, you will need to decide the minimum number of points that students at each achievement level should be able to earn.)

After you have classified each item with a symbol, the D–F cutoff score is found by adding the number of D symbols. Then the C–D cutoff is obtained by adding the number of D and C symbols. The B–C cutoff is the sum of D, C, and B symbols, and the A cutoff is the sum of the D, C, B, and A symbols. To account for negative errors of measurement, you could lower each grade cutoff by one or two points. Such adjustments for error at this stage of grading would make it unnecessary to review borderline cases at a later time.

All grading methods involve subjectivity, and this one requires two main types of subjective decisions. The first type entails the development of explicit expectations for the achievers at each of the letter-grade levels. What is B achievement like and how is it different from C achievement? Good teachers might disagree with one another about how to define these performance standards. The other subjective decision making occurs when items are reviewed to determine the grade category to which each one belongs. Again, good teachers may disagree about whether a "B student" should be able to answer a particular item correctly. Notice that these two types of judgments do not require that subjective decisions be made about individual students. There is no need to decide, for example, whether Jana is a C student or whether Matt could answer a certain question correctly. The judgments required here are about standards and about the particular tasks that students at each level should be expected to do.

Personal Grading Practices Evolve

Since both philosophies and instructional approaches change as curriculum changes, teachers need to be prepared to adjust their grading plans accordingly. With experience in assigning grades, reporting to parents, and observing the impact of grading on learning, many teachers rethink their responses to the philosophical questions enumerated in the "Developing a Grading Philosophy" section. The meanings of the symbols, the characteristics to be judged, the components to include in a grade, and the method used for assigning grades are all issues of value that take on new importance or new meaning as teachers accumulate grading experience and observe the practices of colleagues.

Grading practices also may change as a teacher's instructional approach changes. For example, a teacher who begins experimenting with cooperative learning strategies would start depending more on group projects and presentations for assessment information. The nature of the grading components being used may need to change, as would any grading practices that foster competition among learners.

In short, a teacher's grading practices are likely to evolve slowly over time as his or her grading philosophy changes, as experience in grading accumulates, and as a base of grading data from several classes becomes available. As the nature of the curriculum changes and teachers fine-tune or modify their instructional approaches, the procedures outlined here can be reviewed to adjust inconsistencies in philosophy and practice.

Reference

Oosterhof, A. C. (1987). Obtaining intended weights when combining students' scores. *Educational Measurement: Issues and Practice, 6*(4), 29–37.

Annotated References

The references in this section cover a broad range of topics on grading, as do several other excellent introductory measurement texts. We have chosen to highlight some of the unique or particularly strong parts of these references as an aid to those who seek additional reading.

Carey, L. M. (1988). *Measuring and evaluating school learning.* Newton, MA: Allyn and Bacon. Chapter 13.
The section on designing a gradebook and managing daily records is unique. There also are ample illustrations of the selection and implementation of weights.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th edition). Englewood Cliffs, NJ: Prentice Hall. Chapter 15.
Philosophical issues are discussed in depth and threats to the meaning of grades are considered. The standard deviation and content-based methods are illustrated.

Hills, J. R. (1981). *Measurement and evaluation in the classroom* (2nd edition). Columbus, OH: Bell & Howell. Chapters 14–19.
The chapter on faulty grading practices provides good background for developing a grading philosophy and plan. Broad coverage is also given to reporting methods other than grading.

Oosterhof, A. C. (1990). *Classroom applications of educational measurement.* Columbus, OH: Merrill. Chapters 21–22.
A helpful discussion of the sources of inconsistency in grades is given in one section and a chapter is devoted to weighting procedures for relative grading methods.

Self-Test

1. Which of these statements is most likely to be found in a school's grading policy handbook?
   A. "All teachers will assign grades by grading on the curve."
   B "Grades assigned by teachers are final and may not be appealed."
   C. "Quarter grades must be based on written test scores only."
   D. "The grade of C will be awarded to students whose performance is average compared with their classmates."
2. Which of the following statements indicates that Kathy's B represents her present achievement level compared to an absolute standard?
   A. Kathy is performing well above her peers.
   B. Kathy has shown considerable hard work and has adequate mastery of the primary objectives.
   C. Kathy is the most able student in the class and should have received an A.
   D. Kathy has mastered most of the material taught this grading period.

Use this situation to answer questions 3–5.

Mr. Thompson is a fifth-grade teacher with a class of mixed ability. He has organized his social studies curriculum so that he covers four instructional units per quarter. The following is a list of evaluation data that he collects on each student during each of the four units.

I. 5 homework assignment scores.	III. 1 quiz (after the first week)
II. 1 project	IV. 1 unit test

Mr. Thompson wants his quarter grades to reflect students' achievements at the end of the quarter compared with his absolute standards.

3. Given his grading plan, which information should he incorporate into the composite for the quarter?
   A. IV only
   B. II, III, and IV
   C. II and IV
   D. I, II, III, and IV
4. What additional information is needed to decide how to weight the projects in the final grade?
   A. The amount of time students spent in completing the projects
   B. The amount of variability of the scores within the class on each project
   C. The difficulty level of each project for the class as a whole
   D. The number of unique objectives each project measures compared to other components
5. Which is the best way for Mr. Thompson to grade the projects so that the meaning he wants in his social studies grades is obtained?
   A. Use scoring criteria that are based on content standards.
   B. Try to rank the projects in order from best to worst.
   C. Compare the quality of each student's four projects to look for improvement.
   D. Ask for amount of time spent and amount of help received from others to judge effort.

True–False

6. The objectivity of the standard deviation method for assigning grades makes it superior to the content-based method.
7. Grades that simultaneously incorporate effort, growth, achievement, organization, and ability are less useful than those that incorporate only achievement.
8. One of the advantages of the content-based method over the fixed-percent scale method is that it allows the performance standards for a grade to vary for each component.
9. Some evaluation variables are both reporting variables and grading variables.
10. Homework scores are better grading variables than evaluation variables.
11. If the achievement of a certain objective cannot be measured effectively by a written test, the objective should be excluded from the grading plan.

Answers to Self-Test

1. D. The first three choices are too restrictive or too detailed for most policies. (See Step 1.)
2. D. See Step 2.
3. C. The projects and tests provide information about achievement status at the end of a unit. Homework assignments and quizzes relate to practice and monitoring progress. (See Steps 6–7.)
4. D. The number of objectives covered and the uniqueness of those objectives should be examined when determining component weights. (See Step 7.)
5. A. Components should be scored in a manner consistent with the meaning of the final grade. (See Step 2.)
6. False. The standard deviation method requires subjectivity also (e.g., which grade will be average, how many standard deviations to use to find a cutoff).
7. True. Incorporating more than achievement into a grade distorts the grade's meaning.
8. False. The content-based method allows the cutoff scores to change while keeping the standards constant. (See Content-based Method.)
9. True. Grading variables are a subset of reporting variables, which are in turn a subset of evaluation variables. (See Step 4.)
10. False. Homework is best used to monitor learning and provide practice throughout the instructional unit. (See Step 7.)
11. False. The achievement of such objectives could be assessed by other means—performance assessments, projects, or presentations.